DOCUMENT RESUME

ED 161 893 TM 007 318

AUTHOR Curry, Allen R.; And Others
TITLE Invariance of Rasch Model Ability Parameter Estimates Over Different Collections of Items.
PUB DATE Mar 78
NOTE 82p.; Paper presented at the Annual Meeting of the American Educational Research Association (62nd, Toronto, Ontario, Canada, March 27-31, 1978); Not available in hard copy due to marginal legibility of original document

EDRS PRICE MF-$0.83 Plus Postage. HC Not Available from EDRS.
DESCRIPTORS Complexity Level; Equated Scores; Goodness of Fit; *Item Analysis; *Item Banks; *Mathematical Models; Simulation; Standard Error of Measurement; *Test Construction; Test Items
IDENTIFIERS Birnbaum 2 Parameter Logistic Model; Item Characteristic Curve Theory; Item Discrimination (Tests); Latent Trait Models; *Rasch Model; Test Theory

ABSTRACT

The efficacy of employing subsets of items from a calibrated item pool to estimate the Rasch model person parameters was investigated. Specifically, the degree of invariance of Rasch model ability-parameter estimates was examined across differing collections of simulated items. The ability-parameter estimates were obtained from a simulation of examinee responses to 36 item sets that differed both in fit to the model and in item difficulty. Three examinee groups were formed in order to vary only by shape of the distribution of true ability. The three item-subset characteristics were: (1) percentage of items in the subset that fit the model; (2) difficulty level of the items in the subset; and (3) degree of misfit of those items in the subset that do not fit the model. Each examinee's response to an item was simulated under a general logistic model which takes into account variation in item discrimination. The simulation resulted in an observed score for a subset of items which was based on the examinee's true ability, item difficulty, item discriminations, and a random-error component. (Author/CTM)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED161893

Invariance of Rasch Model Ability Parameter Estimates Over Different Collections of Items

by

Allen R. Curry
Systems Oriented Measurement and Analysis
(SOMA)

W. Louis Bashaw
University of Georgia

R. Robert Rentz
Georgia State University

"PERMISSION TO REPRODUCE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED BY
R. Robert Rentz
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) AND USERS OF THE ERIC SYSTEM."

Paper Presented at the Annual Meeting of the American Educational Rosearch Association, Toronto, Canada, March, 1978

TM007 318

U S DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

Printed in U.S.A.

Invariance of Rasch Model Ability Parameter
Estimates Over Different Collections of Items

Hambleton, Swaminathan, Cook, Eignor, and Gifford (1977) identify three shortcomings of standard testing and measurement methodology, as well as pointing out several problems this methodology has yet to solve. These researchers also note that latent-trait theory and its measurement models hold promise in circumventing these problems and shortcomings. Thus interest in latent trait theory and models is growing among psychometricians. Unfortunately, two major difficulties exist in applying this theory: (a) latent trait models, and in particular the Rasch model, imply that stringent assumptions be met in order for a model to adequately characterize the measurement data (model-data fit); (b) except for the Rasch model, users of latent trait models are experiencing difficulty in estimating model parameters with current mathematical procedures. Thus it would appear that we must "wait" for the other models or attempt to meet the assumptions of the Rasch model. Contrary to this position, however, several researchers indicate that the Rasch model may be robust with respect to violations of assumptions. If this were true, measurement specialists could immediately enjoy the benefits of latent-trait models by employing the Rasch model in situations which allow for relaxation of the assumptions.

The present study involves an exploration of some situations in which the Rasch model assumption of equality of item discriminations is violated. This assumption, which is best termed a condition implied by the model assumptions (c.f., Rentz & Bashaw, 1977), is the most frequently criticized (e.g., Birnbaum, 1968; Hambleton, 1969; Hambleton

& Traub, 1973; Lord, 1968; Ross, 1966). Such criticisms generally follow the form of presenting evidence that items typically vary in discrimination and, therefore, concluding that a model which requires equality is of limited utility. This study attempts to answer Rasch model critics, not by denying that items vary in their discriminations but by demonstrating some conditions which allow for variation and yet allow realization of the outcomes of the Rasch model.

Studies of the robustness of the Rasch model are not uncommon; however, only four studies are known to have addressed the topic of this report--the invariance of the Rasch model person parameter (ability). Tinsley (1971), Whitely and Dawis (1974), Willmott and Fowles (1973), and Wright (1968) have investigated the invariance of ability estimates obtained from different item subsets from a calibrated item pool. Unfortunately, review of these reports indicates (a) conflicting results, and (b) failure to adequately describe and/or control the variables of interest in ability estimation. The first problem may be, at least partially, attributed to variation in the criteria or adequacy of the criteria used to indicate invariance. The second problem, lack of adequate control and description, could be an artifact of using real data. Solutions to these problems may be obtained through data simulation, as employed in this investigation.

In areas that are recently developed and highly complex, such as latent trait theory, simulation serves to simplify the manipulation and control of variables. Data generation provides for simplification of the research design, demands unambiguous definition of variables, and yields explicit control (if a variable is not parameterized in the simulation, it is controlled by exclusion). In general, data simula-

tion may enhance inquiry into complex areas, serving to ease replicability of research and pointing out promising avenues of study.

While several advantages accrue with data simulation, a major concern is the "realism" of the variable values used. In this respect the present study may be unique to Rasch-model research. All values of variables used in this investigation were based upon a review of the Rentz and Bashaw (1975) Rasch analysis of the Anchor Test Study (ATS) data. The ATS (Loret, Seder, Bianchini, & Vale, 1974) was a national test-equating in reading which required administration of nearly 500,000 reading comprehension and vocabulary tests to over 300,000 children in all 50 states. The reanalysis of these data by Rentz and Bashaw produced what is probably the largest single depository of Rasch-model parameter estimates, estimates which are cross-references to traditional item and person statistics. Thus the parameters of this investigation were based not only upon realistic values but also upon values which are verifiably common in their occurrence.

PROBLEM

The aim of this study was to provide information relevant to the efficacy of employing subsets of items from a calibrated item pool to estimate the Rasch model person parameters. In order to provide this information it was necessary to investigate certain characteristics of the ability-estimation situation. Specifically, this study reports an examination of the degree of invariance of Rasch model ability-parameter estimates across differing collections of simulated items. The ability-parameter estimates were obtained from a simulation of

examinee responses to 36 item sets that differed both in fit to the model and in item difficulty.

The research question that served to delineate this study was as follows: do variation in item difficulty and variation in item fit to the Rasch model across sets of items lead to lack of invariant ability estimates for examinees administered these item sets? The intent of this question was to produce a situation in which sets of items of varying characteristics could be compared. Such a comparison would be based upon the degree of invariance of ability estimates produced by each item set. From this comparison it was felt that guidelines for the selection of item subsets from an item pool could be developed so that practical application of subset ability estimation would be facilitated.

Ability Estimation Situation

The ability-estimation situation refers to the conditions under which ability parameters are estimated. Given that one wishes to estimate a group of examinees' abilities by using a subset of items from a calibrated item pool, the estimation situation may be defined as the characteristics of the examinees and the characteristics of the item subset that might influence the estimation process. This study utilized one characteristic to describe examinee group and three characteristics to define an item subset.

Three examinee groups were formed in order to vary only by shape of the distribution of true ability. The three item-subset characteristics were (a) percentage of items in the subset that fit the model, (b) difficulty level of the items in the subset, and (c) degree of misfit of those items in the subset that do not fit the model. These

item-subset characteristics were studied at four, three, and three levels, respectively. Thus, in combination, these characteristics yield 36 item subsets. The encounter of each of the three examinee groups with each of the 36 item subsets defines 108 unique ability-estimation situations.

Item Fit to the Model

Under the general term "model-data fit" may be found a concern for item fit to the Rasch model. Model-data fit refers to the correspondence between what may be termed assumptions of the model and the measurement data; this correspondence is required in order for the model to adequately characterize the measurement situation. In this viewpoint "adequately characterize" implies that correspondence may be a matter of degree. Furthermore, since there are several assumptions to the model, then correspondence may also be a matter of type. These statements are consonant with C. C. Rentz's definition of adequate fit, "a degree and/or type of model-data fit that allows realization of model outcomes (e.g., invariance)" (1975, p. 6).

Those factors commonly termed assumptions, according to Rentz and Bashaw (1977), are actually antecedent conditions deducible from the model assumptions. Of the three commonly described antecedent conditions necessary for model outcomes, only one, equality of item-discrimination values, is of concern in this study. The two remaining conditions, absence of guessing and unidimensionality of the measured trait, were not factors in this data simulation.

Manipulation of item fit to the model was accomplished by varying item-discrimination values. Item-subset fit to the model was varied by altering the percentage of fitting items and the degree of misfit

of non-fitting items within the subset. Three degrees of item misfit were defined: a definition based upon item discrimination values.

Characterizing the Ability Estimation Situation

Characterizing Groups of Subjects

It is conventional in measuring attainment to assume that the trait to be measured is normally distributed within the target population. However, interest in relatively new areas of measurement and learning strategies has brought about involvement with other trait-distribution shapes. For example, if the population of interest is the educationally disadvantaged, a reasonable assumption is that the distribution of most "academic-type" traits would be positively skewed. That is, the preponderance of students would possess very little ability, and as ability level increases, fewer and fewer students would be encountered. In contrast, a successful mastery-learning situation would be expected to evidence the opposite trend, a clustering of students at a relatively high ability level (the mastery or criterion point) with increasingly fewer students at lower ability levels.

Not only is there high interest in these areas, but it would seem that the Rasch model has some unique advantages if applied to these areas. Consider anchoring Rasch abilities to an item pool when measuring the educationally disadvantaged. If it is assumed that the item pool as a whole is highly representative of the trait in question, then fixing ability values relative to the item pool would allow for domain-referenced interpretation of these abilities. While the entire item pool is of interest, it would be possible that many items in the pool would be overly difficult for these students. However, a subset

could probably be found that would be appropriate in difficulty. Administration of this subset would still allow for domain-referenced interpretation of ability values while using an item subset more fitted to the group to be measured.

Utilizing the Rasch model in a mastery-learning situation would involve fixing item-difficulty values relative to the average ability of a group of students who are considered to have mastered the curriculum. Subsequent use of a subset of these items would allow for a quick initial screening of any student whose curriculum mastery was unknown. Those students who are well above the mastery level could proceed to other tasks. Those well below mastery as measured by the item subset would repeat all or portions of the instruction. For those students whose mastery is still questionable, the full item set could be administered. The usage of more items would allow for greater confidence in ability estimates and thus enhance classification of "borderline mastery" students. Considerable time saving, time which might be devoted to instruction, could be realized through such a screening procedure.

If the Rasch model appears useful in situations similar to those of the examples, then knowledge relevant to the attainment of model outcomes in these situations is essential. Noting that fixing model parameter estimates to some standard represents merely a change of measurement unit, it is seen that the two example situations differ from the customary measurement situation only by hypothesized population-distribution shape. However, to the writers' knowledge, no pertinent information is available as to the effect of population-distribution shape on achieving the model outcome of interest in this investigation, invariant ability estimates.

With respect to ability, population-distribution shape is but one of several factors relevant to characterizing groups of examinees. Other factors are average population ability, ability range, variability, and distribution kurtosis. Since we are unsure of the effect of population distribution shape upon ability estimate invariance, it would be necessary to vary this factor while controlling other relevant factors.

Three commonly encountered population-distribution shapes were investigated: normally shaped, positively skewed, and negatively skewed. These three shapes were represented by skewness measures of 0, 1, and -1. In order to prevent confounding of skewness levels with distribution kurtosis, kurtosis was maintained at a mesokurtic level, 3.0. (See Ferguson, 1966, for measures of skewness and kurtosis.)

Specification of a single range of ability values will allow for its control and in addition allow for a determination of an appropriate standard deviation of abilities within the three populations. Equating Reading Tests with the Rasch Model (Rentz & Bashaw, vol. 2, 1975) presents Rasch-model item calibrations and ability estimates for 56 vocabulary and reading comprehension tests. Using these tests for indicators of typical Rasch-model parameter estimates, it is found that the majority of tests yield ability estimates within the natural log-ability interval of approximately -4 to 4. While this interval indicates a rather large range of possible abilities, closer examination of the tests reveals that approximately 50% of the examinees are found on the ability interval of -1 to 1. These findings are not unexpected if the following three factors are considered: (a) a test is commonly built for a particular target group, and as such the test will be centered on that group; (b) Rasch model log abilities will reflect

this centering in that a match of the average item difficulty and person ability will yield an ability estimate of 0; and (c) the Rasch model will produce ability estimates for all test score groups, even those score groups which contain no subjects.

On the basis of the above information it would seem that two ability ranges (8 log units and 2 log units) may be considered appropriate. However, if it is assumed that one purpose of item-subset ability estimation is to measure a relatively homogeneous group of subjects, then the small ability range would be preferred for this data simulation. Thus, the ability interval of -1.0 to +1.0 was adopted for this investigation with the center of this interval taken as the average log ability.

Given this ability interval, an appropriate degree of variability could be represented by a standard deviation of 0.3. This standard deviation appears reasonable since for a normal distribution of abilities the mean ability (0.0) $\pm$ 3 standard deviations covers approximately 99% of the ability interval (-1.0 to +1.0). Even with an unknown distribution, Tchebycheff's inequality (c.f., Hays & Winkler, 1971, p. 167) indicates that approximately 88% of the distribution will be contained between the limits of the mean $\pm$ 3 standard deviations.

Developing three populations of ability values with equal means, ranges, variability, and kurtosis, but differing in skewness would allow for sampling from these populations and subsequent inferences based upon the sample data. One concern in sampling, however, would be that of the number of examinees in each group. In selecting the sample size, it is desired that an adequate number of subjects be obtained so as to "flesh out" the distributions so they are of the desired shape. Further, to prevent confounding of sample size and distribution shape, all groups should contain an equal number of

examinees. The common sample size selected for the three groups of this investigation was 100 examinees. Of interest here is that the number of examinees needed is not influenced by any considerations of sample size required for item calibration. Item calibration is assumed to be completed since item-difficulty estimates on a common scale are a prerequisite to item-subset ability estimation.

Characterizing the Item Pool

Item difficulty. One of the purposes of this investigation was to assess the effect upon ability-estimate invariance of using item subsets of inappropriate difficulty. This purpose implies the existence of an item pool with a difficulty range adequate to select item subsets both appropriate and inappropriate in difficulty. Examination of the item calibration values reported by Rentz and Bashaw (vol. 2, 1975) reveals that intact-test log difficulties are symmetric about 0 with the majority of difficulty values contained on the interval of -2 to 2. Such a difficulty interval is also appropriate for measuring subjects whose log abilities range from approximately -4 to 4 (c.f., Willmott & Fowles, 1973). Now, considering the restricted ability range of this investigation (-1 to 1), an item-pool difficulty interval of -2 to 2 would be quite sufficient for selection of item subsets appropriate and inappropriate in item difficulty.

When items are collected into groups, such as an item pool, then these items may be characterized as to distribution of item difficulty. With respect to distribution, myriad possibilities exist; however, two distributions seem to predominate in practice. According to Rentz and Bashaw:

> models for good tests differ. If tests differ in regards to the distribution of DIFF (item difficulty) then these differences reflect the test authors' opinions of the

> definition of a "good" test for their specific purposes. Some authors will try for a flat, wide distribution of DIFF, while others will try for a peaked, narrow range of DIFF. (1975, vol. 2, p. 9)

The two specific purposes mentioned by Rentz and Bashaw may be viewed as the approximately equally precise measurement of a relatively broad range of subjects and fine discriminations between subjects within a narrow range of ability. These purposes correspond to flat, wide distributions of difficulty and narrow, peaked distributions, respectively. Given a desire to measure as precisely as possible all subjects of this investigation, the wide, flat (rectangular) distribution was seen as appropriate.

The spacing between difficulty values within the item pool can be determined by specifying the number of items in the pool. If we consider a typical test length to be something on the order of 100 items, then it would be reasonable to produce an item pool of similar length. Given 100 items rectangularly distributed on the log-difficulty interval of -2.0 to 2.0, the spacing between difficulty values would be .04 units.

Item discrimination. Typically in Rasch model applications, items may be characterized only by item difficulty; that is, items which fail to fit the model are generally excluded from the item pool. However, if areas for potential applications of the model are to be extended, then tolerable fit deviations, if any, should be determined. Thus, a second item characteristic, item discrimination, was developed within the item pool. In attempting to assess data deviations from the model, Rasch (1960) utilized item-slope values. (The relationship between

item slope and discrimination is given in Appendix A.) These values may be obtained from the plot of the log odds of a correct response to an item for each ability group against the log odds for the group on all items (log odds of a correct response averaged over all items). The Rasch model assumptions imply equality of all such item-slope values for an item pool. However, according to C. C. Rentz, "so long as all items have the same slope, whatever their common value, the Rasch analysis procedure estimates the slope for items that fit the model as 1.0" (1975, p. 86). Therefore, unity was taken as the indicator of item fit to the model.

Researchers, other than Rasch, who have utilized or proposed item discrimination[1] as an index of model fit include Brink (1970), Brooks (1964), Cartledge (1974), Hambleton (1969), Panchapakesan (1969), C. C. Rentz (1975), and Rentz and Bashaw (1975). Of these researchers, Panchapakesan (1969) has indicated an item may be considered fitting if its discrimination is on the interval of .8 to 1.25. C. C. Rentz (1975) and Rentz and Bashaw (1975) support Panchapakesan's findings but indicate the fit interval to be .75 to 1.25.

C. C. Rentz (1975) found that invariance of model-parameter estimates decreases as item-discrimination values depart from the interval of .75 to 1.25. This implies that fit may not be a yes/no proposition

---

[1] Appendix A shows that the slope of the log odds regression line is equivalent to the value of item discrimination. Therefore, the more common term, discrimination, is substituted for the term slope.

but that fit could be a matter of degree. Furthermore, an inference that may be made is that equal deviations from unit imply similar degrees of misfit. That is, an item with a discrimination of 1.5 is equal in model fit to an item of .5 discrimination, while both of these values reflect greater fit than an item with a discrimination of .25.

Considering the regularity of the proposed fit interval ($1 \pm .25$) and noting a lack of other guidance from the literature, it would seem reasonable to group and define misfitting items with similar regularity. One misfit interval so defined could be termed "low misfit" and would be composed of discrimination values on the intervals of .5 to .74 and 1.26 to 1.5. A second degree of misfit, high misfit, would be defined as discrimination values on the intervals of .25 to .49 and 1.51 to 1.75.

Specification of the intervals defining fit and misfit is not sufficient for control of the fit variable that would lead to unambiguous interpretation of the effects of fit. Additional consideration should be paid to the distribution of discrimination values, and the relationship of discrimination to difficulty. In order to prevent confounding of discrimination distribution-shape with levels of fit, shape should remain constant across fit levels. Therefore, the discrimination values were formed to be rectangularly distributed.

The final consideration with respect to discrimination is its relationship to the other item value, item difficulty. Panchapakesan (1969) indicates that, if discrimination values are positively correlated with difficulty, then subject ability will be underestimated. Conversely, if, as difficulty increases, discrimination decreases, then

a person is estimated as being more able than he actually is. In order to eliminate any such bias in the ability estimates obtained in this study, item-discrimination and item-difficulty values were randomly paired.

Characterizing Subsets of Items

As originally conceptualized this study was to employ the following four variables to characterize subsets of items: the difficulty level of the items in the subset, the degree of item misfit within the subset, the percentage of fitting items in the subset, and the length of the subset. However, in combination these variables made for quite a large (over 200) and complicated collection of item subsets to be studied. Some method of reducing the number of subsets was therefore sought. It initially appeared, however, that each of these variables was in need of study and thus none could be summarily excluded. Fortunately, it was discovered that the number of items is not a factor in the actual estimation of ability, but subset length does play a major role in the standard error of an ability estimate. Support for this contention is detailed in Curry (1977).

Subset Length and Standard Error of Measurement

This section provides a review of the standard error of an ability estimate and the rationale for selection of the subset length employed in this investigation.

Wright and Douglas (1975) have shown that the standard error of ability estimate j is given by

$$SEM_j = \sqrt{\frac{1}{\sum_{i=1}^{L} P_{ij}\,(1 - P_{ij})}} \quad , \qquad (1)$$

where P_{ij} = probability of a correct response, and L = number of items used to estimate ability.

From this formulation it becomes apparent that SEM_j is a function of the number of items and the difficulty of these items. Furthermore, the expression $P_{ij}\ (1 - P_{ij})$ will be at a minimum when $P_{ij} = 1 - P_{ij} = .5$. This minimum occurs when the person with ability j has a 50:50 chance of success on an item, and this minimum occurs only when the ability of the person is matched by the difficulty of the item.

Assuming the match of ability and all item difficulties $\sum_{i=1}^{L} P_{ij}\ (1 - P_{ij})$ is equivalent to $L(.5)(.5)$. This match allows SEM_j to be manipulated by variation in L only. That is,

$$SEM_j(\min) = \sqrt{1/L(.25)}. \qquad (2)$$

From equation (2) a table of minimum standard errors can be constructed for item sets of varying length; sets that would yield maximum confidence in ability estimates for those subjects whose ability is matched by all item difficulties. Table 1 provides these minimum SEM_j for item sets of selected lengths.

Noting that a given log ability ± 2 SEM closely approximates the 95% confidence interval for that ability, and since log abilities are commonly found on the interval of -4 to 4, inspection of Table 1 reveals that item sets of 30 or less produce large confidence intervals vis-a-vis the total ability interval. Furthermore, these confidence intervals are the smallest possible. As the ability-difficulty match decreases from optimum, these confidence intervals will become even larger.

Table 1

Minimum Standard Errors of Measurement and Minimum 95% Confidence Interval Coefficients for Item Sets of Varying Length

Item Set Length	Minimum SEM (in log units)	95% Confidence Interval Coefficients
10	.632	± 1.264
20	.447	± 0.894
30	.365	± 0.730
40	.316	± 0.632
50	.283	± 0.566
60	.258	± 0.516
70	.239	± 0.478
80	.224	± 0.448
90	.211	± 0.422
100	.200	± 0.400

In the authors' opinion, application of subset ability estimation employing item subsets of 30 items or less would yield overly imprecise measurement. However, to reduce the 95% confidence interval to less than one log unit would require administration of at least 70 items, far more items than one would expect to encounter in utilizing item subsets. An item subset length of approximately 40 items would seem

a reasonable compromise between overly imprecise measurement and excessive subset length. This investigation employed 40 items as the subset length for all subsets studied.

Item Subset Difficulty

As stated in Characterizing the Item Pool, one of the objectives of this investigation was to assess the effect on ability estimate invariance of measurement with items inappropriate in difficulty. Using inappropriate items could result from inaccurately locating the ability of the target group. For example, assume that we believe the target group to be centered at the log ability level of 1, when in fact the group was located at 0 ability. Following a common measurement practice, we would center our items where we believe the target group to be located, but in reality these items would be "off-target." One obvious effect of using off-target items would be an increase in the standard errors of the ability estimates produced by these items; however, would these ability estimates fail to be invariant?

In defining off-target subsets, mislocations of the target group of one log unit were employed. These mislocations were selected so that off-target subsets corresponding to these mislocations would be appreciably different from appropriate subsets while yielding item probabilities of success that would not reach such extremes that subjects must be omitted due to passing or failing all items in the subset (c.f. Table 2). A mislocation toward high abilities implies using "hard" item subsets, subsets centered at a log difficulty of 1. Furthermore, given 40 items as the subset length and a spacing between difficulty values of .04 units, these subsets would span the log

Table 2

Probabilities of Success on Items of Varying Difficulty for Examinees of Selected Ability

$$P_{ij} = (\exp(\Theta - d_i)/(1 + \exp(\Theta_j - d_i))$$

(Decimals Omitted)

Item Difficulty	Examinee Ability										
	-1.0	-0.8	-0.6	-0.4	-0.2	0.0	0.2	0.4	0.6	0.8	1.0
-1.8	690	731	768	802	832	858	881	900	917	931	943
-1.6	646	690	731	768	802	832	858	881	900	917	931
-1.4	599	646	690	731	768	802	832	858	881	900	917
-1.2	550	599	646	690	731	768	802	832	858	881	900
-1.0	500	550	599	646	690	731	768	802	832	858	881
-0.8	450	500	550	599	646	690	731	768	802	832	858
-0.6	401	450	500	550	599	646	690	731	768	802	832
-0.4	354	401	450	500	550	599	646	690	731	768	802
-0.2	310	354	401	450	500	550	599	646	690	731	768
0.0	269	310	354	401	450	500	550	599	646	690	731
0.2	231	269	310	354	401	450	500	550	599	646	690
0.4	198	231	269	310	354	401	450	500	550	599	646
0.6	168	198	231	269	310	354	401	450	500	550	599
0.8	142	168	198	231	269	310	354	401	450	500	550
1.0	119	142	168	198	231	269	310	354	401	450	500
1.2	100	119	142	168	198	231	269	310	354	401	450
1.4	083	100	119	142	168	198	231	269	310	354	401
1.6	069	083	100	119	142	168	198	231	269	310	354
1.8	057	069	083	100	119	142	168	198	231	269	310

Θ_j = Ability of person j

d_i = Difficulty of item i

difficulty interval of .22 to 1.78. Similarly, underestimating the group's average ability indicates employing "easy" subsets. Such subsets would have an average log difficulty of -1 and would be contained on the difficulty interval of -1.78 to -.22.

In order to provide a contrast to off-target subsets, item subsets appropriate in difficulty should be included in this study. Such subsets would be centered at the difficulty level of 0 and would encompass the interval of -.78 to .78.

Percentage of Fitting Items in the Subset

In a report concerned with ability estimate invariance, Whitely and Dawis (1974) made three comparisons of subsets from a calibrated item pool, a pool not constructed to fit the Rasch model. Two of these comparisons, odd versus even items and two randomly selected subsets, yielded statistically equivalent ability estimates for a group of subjects. However, the third comparison, easy versus hard subsets, did not yield equivalency. These researchers indicate that this lack of ability estimate equivalency may be attributed to poor item fit to the model for the hard subset (57% non-fitting items).

These findings have an intriguing implication for item-subset construction. Namely, it may be necessary to have a core, or minimal number, of fitting items within a subset in order to achieve invariance. This supposition is supported, at least by omission, in that Whitely and Dawis (1974) did not find fault with the easy subset although it contained 23% non-fitting items. Unfortunately, only this slight support from Whitely and Dawis is avaiable; to the knowledge of these authors' no other information exists as to how many items a subset must

contain in order to yield invariant ability estimates.

Due to the paucity of guidance from the literature, it would seem to be advantageous to conduct a general overview of this variable, percentage of fitting items in the subset. Such an overview was undertaken by employing four levels of the variable: 0%, 25%, 50%, and 75%. These levels should be sufficient to introduce differences between subsets so characterized and, if the variable is influential, allow for a rough indication of the percentage of fitting items needed to attain invariance of ability estimates.

With respect to model fit, characterizing a subset merely by percentage of fitting items seems to be incomplete. Can the model-data fit of non-fitting items in the subset be ignored? Can any degree of misfit be tolerated in the non-fitting items? Concerns such as these lead to description of the misfit of non-fitting items in the subset, the topic of the next section of this report.

Characterizing Item Subsets as to Degree of Misfit

In discussing test-fit (or equivalently subset-fit), C. C. Rentz (1975) indicated that a common practice is to gauge the fit of a test to the model by the percentage of fitting items in the test. However, as she notes, R. R. Rentz (1974) has suggested that an index may be insufficient. "For example, a small percentage of 'very bad' items may represent less adequate fit than a larger percentage of 'not too bad' items" (C. C. Rentz, 1975, p. 76). Incorporation of these thoughts into this investigation leads to characterizing item subsets as to the degree of misfit of the subset's non-fitting items.

As noted in the section Characterizing the Item Pool, items may be classified as to degree of misfit; this same item classification was also adapted to describe the degree of misfit of item subsets. Specifically, item subsets containing items all of a low degree of misfit can be described as low misfit subsets. Similarly a high degree of misfit subset would contain only highly misfitting items. (Of course, these descriptors classify the subsets with respect to non-fitting items only; a full description of a subset would include this variable as well as item-difficulty level and percentage of fitting items.) However, having only two categories seems slightly simplistic, for in practice we would also expect to encounter subsets composed of both low and high degrees of misfitting items. Following this expectation, a third level, low and high (mixed) misfit subsets, was included in this investigation.

In order to maintain average discrimination at unity for each subset, item subsets were formed with an equal percentage of items from the appropriate misfit intervals. Table 4 provides the outline for item-subset construction utilized in this investigation.

METHODOLOGY

Generating Examinee True Abilities

As shown previously, this research called for the generation of true abilities for three groups of examinees. These groups were to be randomly selected from three populations with equal true-ability means (0.0) equal standard deviations (0.3) and equal distribution kurtosis (3.0). It was further required that the populations differ in the skewness of the distributions of true ability. The Pearsonian Distribution Package (Bouver & Bargmann, 1974) was selected to perform this generation.

The Pearsonian Distribution Package is a computer program package composed of a sequence of subprograms which evaluates the Pearsonian system of curves in terms of skewness and kurtosis. The application of this package may be explained through the arguments of the control or "master" subprogram, PEARS. The structure of PEARS may be symbolized as

RESULT = PEARS(TPT, BETA1, BETA2, INDEX).

The PEARS subprogram calls different function subroutines (corresponding to different types of Pearsonian distributions) as determined by the values of BETA1 (skewness) and BETA2 (kurtosis). The value of INDEX determines the form of the result. When INDEX is 2, as in this research, the result is the standardized percentage point in the distribution; a percentage point whose probability is TPT.

Random sampling from a population of values may be accomplishing by substituting a random variable for TPT. The random variable should vary uniformly between 0 and 1. After this random sampling, RESULT, which is in standardized form, must be converted into a value which belongs to the distribution with the desired mean and standard deviation. This conversion is

$$\text{value} = \mu + \sigma \cdot \text{RESULT}, \text{ where}$$

μ is the desired mean and σ is the desired standard deviation.

A documented abstract of the main program which utilized the Pearsonian Package is found in Curry (1977). The uniform random variable, which was substituted for TPT, was obtained using the RANF subroutine (Control Data Corporation, FORTRAN Extended Reference Manual).

Constructing Item Subsets

The previously defined variables of percentage of fitting items, level of difficulty, and degree of misfit, when crossclassified, provide specifications for the construction of the item subsets of this investigation. Table 4 provides descriptors of these 36 subsets (the rows of the table) and the discrimination intervals employed to define degrees of model-data fit. The entries in this table are the number of values taken from the indicated discrimination interval.

In forming the item subsets the indicated number of discrimination values was randomly selected from the appropriate discrimination interval. These values were then randomly paired to the difficulty values

Table 3

Number of Item Discrimination Values Used

to Construct 36 Item Subsets

Item Subset* P,M,D	Discrimination Intervals				
	High Misfit (.25-.49)	Low Misfit (.50-.74)	Fit (.75-1.25)	Low Misfit (1.26-1.50)	High Misfit (1.51-1.75)
(0%,Low,Easy)		20		20	
(0%,Low,Moderate)		20		20	
(0%,Low,Hard)		20		20	
(0%,High,Easy)	20				20
(0%,High,Moderate)	20				20
(0%,High,Hard)	20				20
(0%,Mixed,Easy)	10	10		10	10
(0%,Mixed,Moderate)	10	10		10	10
(0%,Mixed,Hard)	10	10		10	10
(25%,Low,Easy)		15	10	15	
(25%,Low,Moderate)		15	10	15	
(25%,Low,Hard)		15	10	15	
(25%,High,Easy)	15		10		15
(25%,High,Moderate)	15		10		15
(25%,High,Hard)	15		10		15
(25%,Mixed,Easy)	7	8	10	7	8
(25%,Mixed,Moderate)	7	8	10	7	8
(25%,Mixed,Hard)	7	8	10	7	8
(50%,Low,Easy)		10	20	10	
(50%,Low,Moderate)		10	20	10	
(50%,Low,Hard)		10	20	10	
(50%,High,Easy)	10		20		10
(50%,High,Moderate)	10		20		10
(50%,High,Hard	10		20		10
(50%,Mixed,Easy)	5	5	20	5	5
(50%,Mixed,Moderate)	5	5	20	5	5
(50%,Mixed,Hard)	5	5	20	5	5
(75%,Low,Easy)		5	30	5	
(75%,Low,Moderate)		5	30	5	
(75%,Low,Hard)		5	30	5	
(75%,High,Easy)	5		30		5
(75%,High,Moderate)	5		30		5
(75%,High,Hard)	5		30		5
(75%,Mixed,Easy)	2	3	30	2	3
(75%,Mixed,Moderate)	2	3	30	2	3
(75%,Mixed,Hard)	2	3	30	2	3

*P-Percentage of Fitting Items, M-Degree of Misfit, D-Difficulty Level

identified by the subset descriptor level of difficulty (e.g., moderate difficulty, 40 difficulty values contained on the interval -.78 to .78). For example, the first row of Table 4 specifies a subset with no fitting items, misfitting items to be taken from the low misfit intervals, and all items to be fromtthe "easy" difficulty interval. Therefore, 20 discrimination values were randomly selected from the interval of .5 to .74, and 20 values were randomly selected from the discrimination interval of 1.26 to 1.5. Each of the resulting 40 values was then randomly paired with one of the 40 difficulty values defining "easy" items.

Item discrimination and difficulty pairings are given in Curry (1977). Descriptive statistics for each of the item subsets employed in this investigation are found appended to this report.

Data Simulation

Given a set of items calibrated under the Rasch model, it is possible to produce a scoring table or a set of ability estimates to be assigned to examinees taking these items. Such a scoring table provides a Rasch-model ability estimate to be associated with each observed score obtained on the item set. Now, if the Rasch model is inadequate to describe the measurement data, then we might expect these ability estimates to be in error. One manner in which the Rasch model may be inadequate is when the measurement data indicates item-discrimination variability. That is, if the factors pertaining to examinees' item responses can best be represented by a model that considers discrimination variation, then the scoring table associations

between the observed scores (sum of these responses) and Rasch model ability estimates may be inaccurate. Such inaccuracies might be sufficient to produce Rasch model ability estimates that lack correspondence to true abilities. Thus, the viewpoint in this data simulation was that, in the presence of variability in item discrimination, a model employing item discrimination should be used to produce item responses. The observed score based on these responses would be used as the referent in a scoring table to select a Rasch model ability estimate, estimates produced under the assumption of equality of discriminations.

To obtain an observed score for an examinee, first we must obtain a set of item-probability values for the items that he encounters in an ability-estimation situation. These probabilities of a correct item response would be based upon the true ability of the examinee, the item difficulty and item discrimination. In logistic form, these probabilities may be obtained from

$$P_{ij} = \exp(a_i(\theta_j - d_i))/(1 + \exp(a_i(\theta_j - d_i))), \qquad (3)$$

where a_i = discrimination of item i,

d_i = difficulty of item i, and

θ_j = the ability parameter or true ability of examinee j.

Conversions of these probabilities into a simulated response vector was accomplished by comparing each item-probability value to a random number which varies uniformly between 0 and 1. The response of person j to item i is labeled "correct" if P_{ij} is greater than or equal to the random number; otherwise, the response is labeled "wrong." The number of correct responses is the subject's total, or observed

score, a score which is based upon the subject's true ability, all of the given item information, and, of course, random error.

The next step is to determine an ability estimate to be associated with this observed score. Wright and Panchapakesan (1969) give the details of an iterative procedure which may be used to produce a scoring table for a set of pre-calibrated items. Solution of the k - 1 ability estimates of the scoring table is based upon

$$j = \sum_{i=1}^{k} (\exp(a_j^* - d_i)/(1 + \exp(a_j^* - d_i))), \qquad (4)$$

where $j = 1, 2, \ldots, k - 1$ = the observed score,

a_j^* = the Rasch log-ability estimate to be associated with each observed score,

d_i = the difficulty of item i, and

k = the number of items in the item set.

To determine the Rasch-model ability estimate for an examinee, all that is required is to obtain from the scoring table the ability estimate associated with that subject's observed score.

The process of data simulation was accomplished through a modification of a computer program developed by R. R. Rentz. The computation of the random numbers used in the production of each examinee's response vector was based upon the power-residual method (c.f., International Business Machines manual, Scientific Subroutine Package program RANDU). A requirement of the computer program was that random-number generation proceed from a "starter" or "seed." A separate random-number starter was used for each of the 36 item subsets of this investigation. These values are contained in an appendix. An abstract

of the program used in the simulation of the data is also contained in an appendix.

Defining a Measure of Invariance and Data Analysis

The encounter of a group of examinees with a set of pre-calibrated items has been defined as an ability-estimation situation. If from one encounter to another the groups change, or the items change, or if both groups and items change, then we have a different ability-estimation situation. In order to compare these ability-estimation situations with respect to ability-estimate invariance, a measure of invariance is needed.

In choosing to perform data simulation for this investigation, one consideration was that, in simulation, the input ability values are in fact true ability parameter values. Furthermore, given these true ability parameters, it is possible to gauge the accuracy of an estimate of an individual's ability by determining the correspondence of that estimate to his known true ability. If we estimate abilities for a group of subjects, we may describe the accuracy of the estimates for the group by averaging the correspondences of each group member's estimate and his true ability. This average ability estimate accuracy may be viewed as a measure of the ability-estimate invariance for the particular ability-estimation situation defined by the group's encounter with the set of items used to obtain ability estimates. In addition, average ability estimate accuracy may be used as a basis for comparisons of ability-estimate situations.

To accomplish these comparisons it is necessary to fully define the accuracy measure for an individual, i.e., the type of correspondence between an individual's ability estimate and his true ability must be specified. C. C. Rentz (1975) indicates that comparisons of estimates and parameters may be made by taking either the standardized or the absolute difference between the parameter and the estimate. Employing standardized differences would seem to imply that for each subject we must obtain the mean difference between each estimate and his true ability, the standard deviation of these differences, and then transform each of these differences into a z-score. As the number of subjects and/or the number of ability estimates for each subject increases, the computational effort for this procedure would become quite great. Furthermore, if having collected these standardized differences, we wished to compare additional ability-estimation situations, it would be necessary to repeat the entire standardizing process while including these new ability estimates. Employing absolute differences between parameters and estimates would circumvent the problems associated with using standardized differences. C. C. Rentz (1975) notes that

> if analysis of variance were applied to. . .(either standardized or absolute) difference data, lack of a significant effect between sets of estimates could be interpreted as indicating the presence of invariance. Such a procedure would allow judgement of invariance when the variation in the estimates was of no greater magnitude than would be expected by chance. (p. 63)

Thus, it would seem appropriate to employ absolute differences as a measure of invariance and then to isolate components of ability-

estimation situations using this measure in an analysis of variance. Within the context of this investigation, a repeated-measure analysis of variance would be indicated with one between-subjects variable, true-ability distribution shape, and three within-subjects variables—item-subset difficulty, percentage-of-fitting items, and subset degree of misfit.

RESULTS

Data Simulation

Complete details of the rationale and method of the data simulation were given previously. However, to recapitulate, each examinee's response to an item was simulated under a general logistic model which takes into account variation in item discrimination. The simulation resulted in an observed score (number of correct responses) for a subset of items which was based on the examinees' true ability, item difficulty, item discriminations, and a random-error component. These observed scores were then used to estimate each examinee's ability under the Rasch model. If the Rasch model is insufficient to account for variation in item discrimination, then this insufficiency should be reflected in the ability estimates produced under the model. Such insufficiency could be detected by employing a measure of the correspondence between the examinees' Rasch-model ability estimates and the examinee's true abilities.

Data-simulation summary statistics are reported for these three types of population true-ability distribution shapes—normal distribution, positively-skewed distribution, and negatively-skewed distribu-

ion. Tables 4, 5, and 6 provide summaries for three variables: observed score, Rasch model ability estimate, $\hat{\theta}$, and the absolute difference between ability estimate and true ability, $|\hat{\theta} - \theta|$. Table 7 displays the product-moment correlations for true ability and the 36 ability-estimate sets of this study.

In order to ascertain when $|\hat{\theta} - \theta|$ reflects a lack of invariance it would be necessary to know what values of this measure should be anticipated when there is invariance. Within the context of this study, we would expect invariance to occur when all items of a subset conform to the Rasch model. As noted in previous sections this conformity is attained when all item discriminations are equal.

The "perfect fit" data simulation, summarized in Table 8, was made under the simulation method described above with the exception that all item discriminations were set to unity. Thus, the reported values of $|\hat{\theta} - \theta|$ should approximate the values anticipated under invariance when random error is allowed to operate.

Simulation Results

The perfect-fit simulation, Table 8, shows that moderate-difficulty items produce lower values of $|\hat{\theta} - \theta|$, indicating a greater degree of invariance. Hard items seem to reflect middle levels of invariance, while easy items are considerably less invariant. The ordering is the same for all three subject groups; however, the magnitude of the dependent variable differs, although without statistical significance, from group to group. The correlations of $\hat{\theta}$ and θ also reflect the moderate-hard-easy ordering; however, the correlations for the positive and negative skew groups are much reduced from the normal groups' correlations. This reduction is reflective of the cor-

Table 4

Simulation Results Summary for Normal

Distribution of True Ability

True Ability (θ)

$\bar{X}$ = -0.066 sd = .535 min = -0.967 max = 0.952

Item Subset* P, M, D	Observed Score $\bar{X}$	sd	min	max	Rasch Model $\hat{\theta}$ $\bar{X}$	sd	min.	max	$\lvert\hat{\theta} - \theta\rvert$ $\bar{X}$	sd	min	max
0%, Low, Easy	27.350	4.922	13	38	-0.120	0.654	-1.768	2.038	0.350	0.255	0.017	1.271
0%, Low, Moderate	19.600	5.492	5	32	-0.046	0.622	-2.025	1.449	0.255	0.192	0.001	1.073
0%, Low, Hard	11.380	4.709	2	26	-0.053	0.676	-2.039	1.651	0.311	0.301	0.001	1.633
0%, High, Easy	26.800	4.524	16	35	-0.212	0.554	-1.427	1.024	0.329	0.243	0.002	1.260
0%, High, Moderate	19.880	5.498	7	30	-0.017	0.617	-1.619	1.151	0.277	0.184	0.006	0.726
0%, High, Hard	12.260	4.764	4	26	0.081	0.612	-1.281	1.651	0.349	0.245	0.004	1.041
0%, Mixed, Easy	27.270	4.677	15	35	-0.146	0.587	-1.537	1.024	0.315	0.235	0.004	0.906
0%, Mixed, Moderate	19.350	5.237	9	31	-0.073	0.582	-1.295	1.295	0.234	0.193	0.004	0.893
0%, Mixed, Hard	11.620	4.711	0	21	-0.018	0.664	-2.763	1.105	0.323	0.301	0.005	2.194
25%, Low, Easy	27.230	5.245	16	39	-0.122	0.716	-1.427	2.763	0.317	0.262	0.017	1.852
25%, Low, Moderate	19.090	6.270	7	31	-0.105	0.707	-1.619	1.295	0.304	0.227	0.005	0.917
25%, Low, Hard	11.620	4.383	3	21	0.001	0.581	-1.601	1.105	0.259	0.193	0.193	0.836
25%, High, Easy	27.450	4.314	15	37	-0.125	0.567	-1.537	1.601	0.309	0.253	0.002	1.260
25%, High, Moderate	19.270	5.124	10	31	-0.080	0.570	-1.151	1.295	0.246	0.173	0.007	0.845
25%, High, Hard	11.990	4.957	3	23	0.031	0.664	-1.601	1.318	0.337	0.257	0.005	1.120
25%, Mixed, Easy	27.640	4.719	16	37	-0.088	0.627	-1.427	1.601	0.326	0.261	0.006	1.554
25%, Mixed, Moderate	19.610	4.932	7	30	-0.045	0.546	-1.619	1.151	0.243	0.181	0.003	0.905
25%, Mixed, Hard	11.820	4.791	3	24	0.013	0.646	-1.601	1.426	0.304	0.226	0.000	1.125
50%, Low, Easy	27.960	5.179	18	38	-0.018	0.728	-1.211	2.038	0.297	0.232	0.006	1.155
50%, Low, Moderate	19.170	5.498	7	30	-0.096	0.616	-1.619	1.151	0.240	0.200	0.004	0.979
50%, Low, Hard	10.830	4.652	1	23	-0.136	0.698	-2.763	1.318	0.345	0.338	0.005	2.394
50%, High, Easy	27.690	4.798	16	37	-0.081	0.634	-1.427	1.601	0.308	0.243	0.000	1.152
50%, High, Moderate	19.600	5.612	9	32	-0.046	0.629	-1.295	1.449	0.269	0.225	0.001	1.255
50%, High, Hard	11.790	4.891	1	25	-0.001	0.691	-2.763	1.537	0.348	0.296	0.005	2.044
50%, Mixed, Easy	27.460	4.802	18	36	-0.113	0.619	-1.211	1.281	0.345	0.236	0.006	1.037
50%, Mixed, Moderate	19.070	5.934	8	31	-0.105	0.666	-1.449	1.295	0.296	0.193	0.018	0.944
50%, Mixed, Hard	11.370	4.792	2	23	-0.062	0.691	-2.039	1.318	0.343	0.246	0.004	1.087

Table 4 continued

Item Subset* P, H, D	Observed Score $\bar{X}$	sd	min	max	Rasch Model $\hat{\theta}$ $\bar{X}$	sd	min	max	$\lvert\hat{\theta} - \theta\rvert$ $\bar{X}$	sd	min	max
75%, Low, Easy	28.090	5.027	15	38	-0.017	0.681	-1.537	2.038	0.284	0.222	0.002	1.155
75%, Low, Moderate	19.140	5.956	6	29	-0.104	0.671	-1.808	1.017	0.307	0.223	0.007	1.005
75%, Low, Hard	11.270	4.830	3	23	-0.065	0.660	-1.601	1.318	0.303	0.233	0.003	1.031
75%, High, Easy	27.530	4.779	16	37	-0.103	0.627	-1.427	1.601	0.303	0.262	0.001	1.262
75%, High, Moderate	19.350	5.321	10	32	-0.072	0.592	-1.151	1.449	0.277	0.219	0.007	1.123
75%, High, Hard	11.020	4.483	1	24	-0.104	0.677	-2.763	1.426	0.315	0.299	0.000	1.904
75%, Mixed, Easy	28.140	4.948	13	39	-0.013	0.687	-1.768	2.763	0.304	0.286	0.005	1.852
75%, Mixed, Moderate	19.380	5.399	9	31	-0.070	0.601	-1.295	1.295	0.264	0.200	0.004	1.056
75%, Mixed, Hard	11.140	4.383	3	23	-0.071	0.609	-1.601	1.318	0.301	0.239	0.002	1.013

*P = Percentage of Fitting Items, H = Degree of Misfit, D = Difficulty Level

Table 5

Simulation Results Summary for Positively Skewed Distribution of True Ability

True Ability (θ)

$\bar{X}$ = -0.017 sd = 0.290 min = 0.300 max = 0.807

Item Subset* P, H, D	Observed Score $\bar{X}$	sd	min	max	Rasch Model $\hat{\theta}$ $\bar{X}$	sd	min	max	$\lvert\hat{\theta} - \theta\rvert$ $\bar{X}$	sd	min	max
0%, Low, Easy	27.990	3.320	19	36	-0.078	0.435	-1.105	1.281	0.305	0.230	0.010	1.168
0%, Low, Moderate	20.220	3.280	11	28	0.024	0.354	-1.017	0.889	0.238	.178	.006	.717
0%, Low, Hard	11.500	3.735	3	23	0.000	.502	-1.601	1.318	.278	.258	.001	1.301
0%, High, Easy	27.560	3.382	20	37	- .130	.447	-1.000	1.601	.326	.256	.002	1.837
0%, High, Moderate	20.290	3.955	13	30	.034	.431	- .768	1.151	.272	.201	.017	.933
0%, High, Hard	11.910	3.559	4	23	.063	.456	-1.281	1.318	.305	.246	.005	.983
0%, Mixed, Easy	28.060	3.510	21	36	- .062	.468	- .895	1.281	.315	.231	.019	1.144
0%, Mixed, Moderate	19.560	3.919	11	30	- .047	.430	-1.017	1.151	.246	.187	.002	.769
0%, Mixed, Hard	11.680	3.910	4	23	.022	.518	-1.281	1.318	.336	.272	.000	1.067
25%, Low, Easy	28.090	3.491	20	36	- .056	.479	-1.000	1.281	.306	.221	.000	1.104
25%, Low, Moderate	19.270	4.206	10	30	- .077	.463	-1.151	1.151	.288	.214	.008	.854
25%, Low, Hard	11.620	3.284	4	19	.027	.434	-1.281	.895	.268	.194	.005	.982
25%, High, Easy	28.110	2.930	21	36	- .067	.396	- .895	1.281	.278	.254	.000	1.242
25%, High, Moderate	19.670	4.045	13	31	- .032	.443	- .768	1.295	.254	.183	.005	.933
25%, High, Hard	11.960	3.784	4	21	.066	.479	-1.281	1.105	.283	.237	.001	1.065
25%, Mixed, Easy	28.610	3.296	21	38	.010	.472	- .895	2.038	.294	.277	.000	1.869
25%, Mixed, Moderate	20.100	3.480	13	29	.012	.376	- .768	1.017	.222	.185	.005	.793
25%, Mixed, Hard	12.040	3.750	5	22	.075	.483	-1.025	1.211	.312	.203	.003	.838
50%, Low, Easy	28.480	3.480	19	36	- .009	.463	-1.105	1.281	.312	.211	.000	.899
50%, Low, Moderate	19.830	3.944	7	29	- .021	.438	-1.619	1.017	.272	.225	.006	1.345
50%, Low, Hard	10.920	3.395	4	21	- .070	.460	-1.281	1.105	.317	.253	.000	1.472
50%, High, Easy	28.280	3.297	21	37	- .032	.473	- .895	1.601	.284	.245	.000	.955
50%, High, Moderate	19.710	3.875	10	29	- .031	.421	-1.151	1.017	.280	.215	.002	.934
50%, High, Hard	11.820	3.406	6	24	.056	.429	- .808	1.426	.280	.222	.001	.922
50%, Mixed, Easy	28.350	3.377	21	36	- .027	.455	- .895	1.281	.336	.230	.005	.952
50%, Mixed, Moderate	19.310	4.099	12	31	- .072	.448	- .889	1.295	.259	.190	.004	.869
50%, Mixed, Hard	11.300	3.791	5	23	-.027	.507	-1.025	1.318	.322	.249	.000	.944

Table 5 continued

Item Subset* P, H, D	Observed Score $\bar{X}$	sd	min	max	Rasch Model $\hat{\theta}$ $\bar{X}$	sd	min	max	$\lvert\hat{\theta} - \theta\rvert$ $\bar{X}$	sd	min	max
75%, Low, Easy	28.670	3.367	22	37	.019	.475	- .789	1.601	.275	.247	.003	1.130
75%, Low, Moderate	20.030	4.029	12	33	.004	.443	- .889	1.619	.289	.213	.007	.990
75%, Low, Hard	11.320	3.741	5	21	- .021	.491	-1.025	1.105	.303	.225	.000	.857
75%, High, Easy	28.300	3.298	21	38	- .030	.474	- .895	2.038	.323	.267	.000	1.333
75%, High, Moderate	19.680	3.985	9	32	- .033	.440	-1.295	1.449	.285	.218	.005	1.007
75%, High, Hard	11.010	3.298	4	19	- .057	.454	-1.281	.895	.313	.225	.005	1.023
75%, Mixed, Easy	28.270	3.698	19	38	.033	.519	-1.105	2.038	.303	.268	.006	1.331
75%, Mixed, Moderate	19.900	3.914	11	31	- .010	.427	-1.017	1.295	.259	.202	.004	1.038
75%, Mixed, Hard	10.910	3.835	3	20	- .084	.523	-1.601	1.000	.323	.258	.007	1.364

*P = Percentage of Fitting Items, H = Degree of Misfit, D = Difficulty Level

Table 6

Simulation Results Summary for Negatively

Skewed Distribution of True Ability

True Ability (θ)

$\bar{X}$ = 0.039 sd = 0.292 min = -0.835 max = 0.300

Item Subset* P, M, D	Observed Score $\bar{X}$	sd	min	max	Rasch Model $\hat{\theta}$ $\bar{X}$	sd	min	max	$\lvert\hat{\theta} - \theta\rvert$ $\bar{X}$	sd	min	max
0%, Low, Easy	28.640	3.526	16	36	0.009	0.459	-1.427	1.281	0.295	0.243	0.003	1.172
0%, Low, Moderate	20.560	4.098	6	28	.056	.455	-1.808	0.889	.258	.199	.015	1.115
0%, Low, Hard	11.820	3.089	5	20	.057	.411	-1.025	1.000	.269	.228	.013	.922
0%, High, Easy	28.120	3.406	16	37	- .063	.438	-1.427	1.601	.304	.236	.003	1.414
0%, High, Moderate	20.740	3.858	11	29	.079	.420	-1.017	1.017	.251	.190	.000	.801
0%, High, Hard	12.330	3.450	5	20	.115	.450	-1.025	1.000	.285	.204	.000	.800
0%, Mixed, Easy	28.460	3.512	19	35	- .013	.460	-1.105	1.024	.301	.226	.000	1.120
0%, Mixed, Moderate	20.080	4.146	6	32	.005	.467	-1.808	1.449	.228	.226	.002	1.215
0%, Mixed, Hard	12.060	3.730	2	21	.066	.533	-2.039	1.105	.343	.315	.004	1.553
25%, Low, Easy	28.490	3.295	19	36	- .011	.445	-1.105	1.281	.325	.219	.003	1.008
25%, Low, Moderate	19.930	4.108	8	28	- .001	.453	-1.449	.889	.280	.190	.002	.718
25%, Low, Hard	12.000	3.005	4	21	.080	.404	-1.281	1.105	.258	.173	.001	.834
25%, High, Easy	28.120	3.367	19	36	- .059	.447	-1.105	0.281	.298	.244	.000	1.256
25%, High, Moderate	20.480	4.019	11	31	.052	.440	-1.017	1.295	.261	.190	.008	1.044
25%, High, Hard	12.300	3.694	3	21	.106	.487	-1.601	1.105	.309	.253	.001	1.072
25%, Mixed, Easy	28.810	3.578	20	37	.044	.509	-1.000	1.601	.304	.279	.003	1.734
25%, Mixed, Moderate	20.620	3.789	12	30	.067	.411	- .889	1.151	.239	.180	.002	.856
25%, Mixed, Hard	12.500	3.433	5	20	.138	.444	-1.025	1.000	.293	.197	.011	.813
50%, Low, Easy	28.780	3.480	21	37	.032	.472	- .895	1.601	.324	.235	.002	1.356
50%, Low, Moderate	20.510	4.110	8	29	.057	.458	-1.449	1.017	.250	.212	.000	1.211
50%, Low, Hard	11.140	3.143	3	20	- .037	.439	-1.601	1.000	.314	.272	.000	1.561
50%, High, Easy	28.670	3.330	19	36	.012	.444	-1.105	1.281	.279	.226	.003	1.156
50%, High, Moderate	20.410	3.854	9	29	.044	.422	-1.295	1.017	.265	.216	.007	1.022
50%, High, Hard	12.380	3.219	4	21	.128	.412	-1.281	1.105	.278	.217	.004	1.000
50%, Mixed, Easy	28.780	3.617	18	36	.032	.478	-1.211	1.281	.318	.227	.003	.990
50%, Mixed, Moderate	19.930	3.585	11	30	-.007	.389	-1.017	1.151	.260	.190	.008	.853
50%, Mixed, Hard	11.830	3.599	3	18	.040	.506	-1.601	.789	.321	.228	.005	.908

Table 6 continued

Item Subset* P, M, D	Observed Score $\bar{X}$	sd	min	max	Rasch Model $\hat{\theta}$ $\bar{X}$	sd	min	max	$\lvert\hat{\theta} - \theta\rvert$ $\bar{X}$	sd	min	max
75%, Low, Easy	29.070	2.982	21	36	.061	.414	- .895	1.291	.261	.212	.003	1.110
75%, Low, Moderate	20.250	4.210	10	30	.027	.462	-1.151	1.151	.285	.209	.002	.918
75%, Low, Hard	11.660	3.412	4	19	.026	.464	-1.281	.895	.314	.245	.011	1.004
75%, High, Easy	28.600	3.542	20	38	.010	.489	-1.000	2.038	.304	.264	.003	1.830
75%, High, Moderate	20.300	3.815	11	30	.033	.417	-1.017	1.151	.272	.222	.000	.942
75%, High, Hard	11.500	3.512	2	19	- .005	.514	-2.039	.895	.316	.247	.005	1.204
75%, Mixed, Easy	29.010	3.762	17	36	.066	.501	-1.318	1.281	.279	.260	.003	1.138
75%, Mixed, Moderate	20.570	4.258	6	30	.058	.474	-1.808	1.515	.278	.212	.003	.989
75%, Mixed, Hard	11.240	3.300	4	20	- .027	.457	-1.281	1.000	.299	.231	.013	1.153

*P = Percentage of Fitting Items, M = Degree of Misfit, D = Difficulty Level

Table 7

Correlations of True Ability and Rasch Models

Ability Estimates for Three Subject Groups

Item Subset* P, M, D	Correlations: Normal Distribution	Correlations: Positive Skew	Correlations: Negative Skew	Regression Line Slope: Normal Distribution	Regression Line Slope: Positive Skew	Regression Line Slope: Negative Skew
0%, Low, Easy	.76	.52	.56	.62	.34	.36
0%, Low, Moderate	.86	.60	.70	.74	.49	.45
0%, Low, Hard	.77	.66	.54	.61	.38	.38
0%, High, Easy	.75	.48	.54	.73	.31	.36
0%, High, Moderate	.85	.63	.67	.73	.42	.46
0%, High, Hard	.76	.54	.65	.67	.34	.42
0%, Mixed, Easy	.77	.56	.58	.70	.35	.37
0%, Mixed, Moderate	.86	.70	.74	.79	.47	.46
0%, Mixed, Hard	.75	.55	.49	.61	.31	.27
25%, Low, Easy	.82	.62	.30	.62	.37	.33
25%, Low, Moderate	.85	.64	.67	.64	.40	.43
25%, Low, Hard	.84	.65	.65	.77	.44	.47
25%, High, Easy	.74	.44	.56	.70	.32	.36
25%, High, Moderate	.85	.71	.68	.80	.46	.45
25%, High, Hard	.78	.66	.59	.63	.40	.35
25%, Mixed, Easy	.75	.52	.58	.64	.32	.33
25%, Mixed, Moderate	.84	.65	.59	.83	.50	.49
25%, Mixed, Hard	.82	.67	.64	.68	.40	.42
50%, Low, Easy	.87	.58	.53	.64	.36	.33
50%, Low, Moderate	.86	.59	.70	.75	.39	.45
50%, Low, Hard	.73	.50	.43	.56	.31	.29
50%, High, Easy	.79	.59	.59	.66	.36	.39
50%, High, Moderate	.83	.56	.59	.71	.38	.41
50%, High, Hard	.76	.59	.37	.59	.39	.41
50%, Mixed, Easy	.75	.47	.57	.65	.30	.35
50%, Mixed, Moderate	.85	.71	.59	.68	.46	.44
50%, Mixed, Hard	.79	.59	.63	.61	.34	.36
75%, Low, Easy	.85	.63	.59	.67	.38	.42
75%, Low, Moderate	.83	.59	.64	.66	.38	.41
75%, Low, Hard	.81	.64	.52	.66	.38	.33
75%, High, Easy	.77	.48	.37	.66	.29	.34
75%, High, Moderate	.81	.58	.53	.73	.38	.38
75%, High, Hard	.77	.54	.63	.61	.35	.36
75%, Mixed, Easy	.80	.64	.65	.62	.36	.38
75%, Mixed, Moderate	.84	.64	.69	.74	.43	.42
75%, Mixed, Hard	.78	.63	.58	.69	.35	.37

*P = Percentage of Fitting Items, M = Degree of Misfit, D = Difficulty Level

Table 8

Descriptive Statistics For Perfect Fit Simulation

Normal Distribution of

True Ability (θ) $\bar{X}$ = -0.066 sd = .535 minimum = -0.967 maximum = 0.952

Item Subset Difficulty	Observed Score $\bar{X}$	sd	min	max	Rasch Model $\hat{\theta}$ $\bar{X}$	sd	min	max	$\lvert\hat{\theta} - \theta\rvert$ $\bar{X}$	sd	min	max	$r(\hat{\theta},\theta)$
Easy	27.96	4.097	14	38	-0.028	0.698	-1.651	2.038	0.342	0.269	0.002	1.377	0.78
Moderate	19.49	5.613	4	33	-0.059	0.649	-2.281	1.619	0.254	0.208	0.004	1.329	0.86
Hard	10.98	5.103	2	25	-0.124	0.731	-2.039	1.537	0.297	0.264	0.000	1.195	0.85

Positive Skewed Distribution of

True Ability (θ) $\bar{X}$ = -0.017 sd = 0.290 minimum = -0.300 maximum = 0.807

Item Subset Difficulty	Observed Score $\bar{X}$	sd	min	max	Rasch Model $\hat{\theta}$ $\bar{X}$	sd	min	max	$\lvert\hat{\theta} - \theta\rvert$ $\bar{X}$	sd	min	max	$r(\hat{\theta},\theta)$
Easy	28.78	3.295	19	36	0.030	0.458	-1.105	1.281	0.308	0.216	0.017	0.917	0.57
Moderate	20.17	3.438	12	29	0.018	0.371	-0.889	1.017	0.258	0.171	0.006	0.754	0.59
Hard	11.08	3.754	4	23	- .037	0.509	-1.281	1.318	0.277	0.247	0.004	1.051	0.70

Negative Skewed Distribution of

True Ability (θ) $\bar{X}$ = 0.039 sd = 0.292 minimum = -0.835 maximum = 0.300

Item Subset Difficulty	Observed Score $\bar{X}$	sd	min	max	Rasch Model $\hat{\theta}$ $\bar{X}$	sd	min	max	$\lvert\hat{\theta} - \theta\rvert$ $\bar{X}$	sd	min	max	$r(\hat{\theta},\theta)$
Easy	29.11	3.524	17	36	0.075	0.473	-1.318	0.281	0.299	0.242	0.004	1.001	0.58
Moderate	20.48	4.016	8	26	0.047	0.443	-1.449	0.651	0.247	0.172	0.006	0.756	0.74
Hard	11.53	3.332	3	21	0.013	0.453	-1.601	1.105	0.269	0.236	0.006	1.140	0.61

relation coefficient's sensitivity to a restriction in range for the variables of these two groups.

The simulation results when item discriminations are variant, Tables 4,5, and 6, show similarity to the perfect fit data. However, in a few instances large values for $|\hat{\theta} - \theta|$ are encountered. In each of these instances large differences between the ability estimate and the true ability are associated with examinees whose observed scores are near zero or near the maximum, 39, but these examinees do not have the minimum or maximum true abilities. Therefore, the large values of $|\hat{\theta} - \theta|$ seem to be due to the random error component of the data simulation.

Analysis

The measure of invariance, $|\hat{\theta} - \theta|$, was employed as the dependent variable in a repeated-measures analysis of variance. This analysis involved one between-subjects variable, true-ability distribution shape, and three within-subjects variables, percentage-of-fitting items, degree of item misfit, and difficulty level of the items in the subset. The data of this analysis may be viewed as the result of administration of 36 differing subsets of items to three groups of examinees.

As indicated in Table 9, four sources of variation were found to be statistically significant--examinee group-distribution shape (G), subset difficulty level (D), the interaction of difficulty and degree of misfit (MD), and the interaction of percentage-of-fitting items and difficulty level (PD). Cell means for the significant sources of variation were presented in Table 10.

The cell means for difficulty level reflect the same ordering as do the perfect-fit data. That is, items appropriate in difficulty

Table 9

Analysis of Variance Summary Table

for Dependent Variable

Source	df	Error Term	Mean Square	F
Group Distribution Shape (G)	2	R(G)	0.1951	3.65*
Percentage Fitting Items (P)	3	PR(G)	0.0792	1.40
Misfit Degree (M)	2	MR(G)	0.0160	0.36
Difficulty Level (D)	2	DR(G)	2.2544	34.98***
GP	6	PR(G)	0.0292	0.52
GM	4	MR(G)	0.0100	0.23
PM	6	PMR(G)	0.0656	1.27
GD	4	DR(G)	0.0166	0.26
PD	6	PDR(G)	0.1314	2.54**
MD	4	MDR(G)	0.1513	2.91**
GPM	12	PMR(G)	0.0292	0.57
GPD	12	PDR(G)	0.0196	0.38
GMD	8	MDR(G)	0.0353	0.68
PMD	12	PMDR(G)	0.0978	1.73
GPMD	24	PMDR(G)	0.0132	0.23
Replicates (G)	297		0.0534	
PR(G)	891		0.0565	
MR(G)	594		0.0440	
DR(G)	594		0.0645	
PMR(G)	1782		0.0515	
PDR(G)	1782		0.0517	
MDR(G)	1188		0.0521	
PMDR(G)	3564		0.0566	

*p < .05

**p < .01

***p < .001

Table 10

Cell Means for Significant Sources of Variation

Group		
Normal Distribution	Positive Skew	Negative Skew
0.3009	0.2907	0.2866

Difficulty Level		
Easy Items	Moderate Items	Hard Items
0.3068	0.2638	0.3055

Percentage Fitting X Difficulty Level

	Easy	Moderate	Hard
0% Fitting	0.3156	0.2509	0.3109
25% Fitting	0.3064	0.2596	0.2914
50% Fitting	0.3126	0.2655	0.3186
75% Fitting	0.2929	0.2797	0.3095

Misfit Degree X Difficulty Level

	Easy	Moderate	Hard
Low Misfit	0.3042	0.2720	0.2948
High Misfit	0.3047	0.2675	0.3097
Mixed Misfit	0.3117	0.2521	0.3182

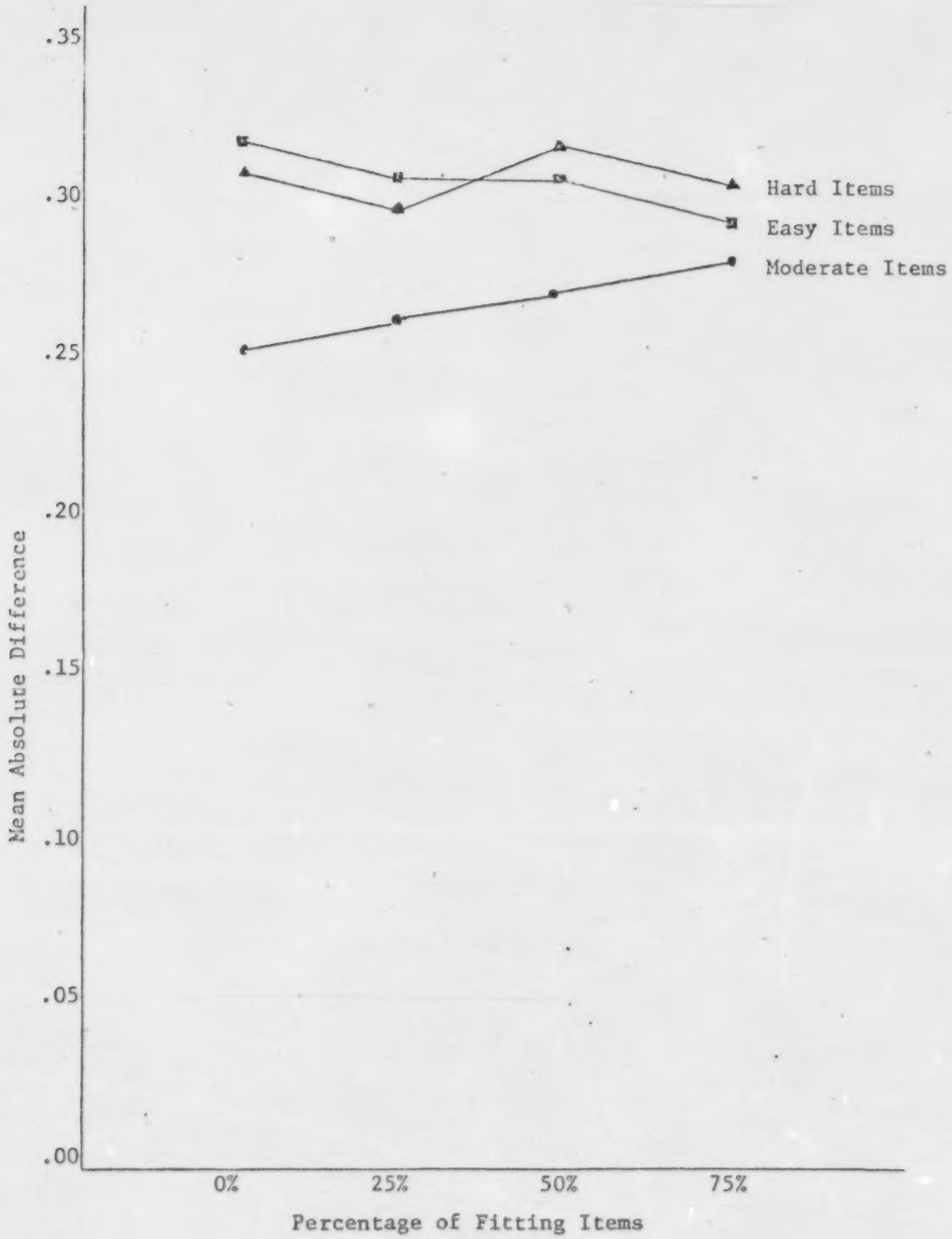

Figure 1. Cell Means for the Interaction of Percentage of Fitting Items and Difficulty

for the ability of the measured examinees yield a higher degree of invariance (lower values of $|\hat{\theta} - \theta|$). Similarly, the perfect-fit data, averaged over difficulty level, and the simulation results show the ordering with respect to invariance to be the negative-skew distribution, followed by positive-skew and normal distributions.

The graph of the interaction of percentage-of-fitting items and difficulty shows an interesting trend (Figure 1). While moderate items still seem preferential, the ordering for these items across the levels of the percentage-of-fitting-items factor is unexpected. It would be anticipated that, as the percentage-of-fitting items increases, we would have greater conformity to the Rasch model and thus higher degrees of invariance. Nonetheless, the opposite trend is found for moderate items.

The graph of misfit degree by difficulty-level interaction-cell means is given in Figure 2. These means again indicate that moderate items yield a higher degree of invariance. However, with the exception of the low degree of misfit items, the perfect-fit ordering is reversed for easy and hard items.

INVARIANCE OF RASCH MODEL ABILITY ESTIMATES AND ADDITIONAL SIMULATIONS

Invariance of Rasch Model Ability Estimates

As noted earlier, the data simulation results were unexpected. The 36 item subsets were ordered from low to high by theoretical conformity to the Rasch model; however, the results do not reflect this theoretical ordering. For example, Figure 1, which presents the graph of the percentage of fitting items by difficulty level interaction cell

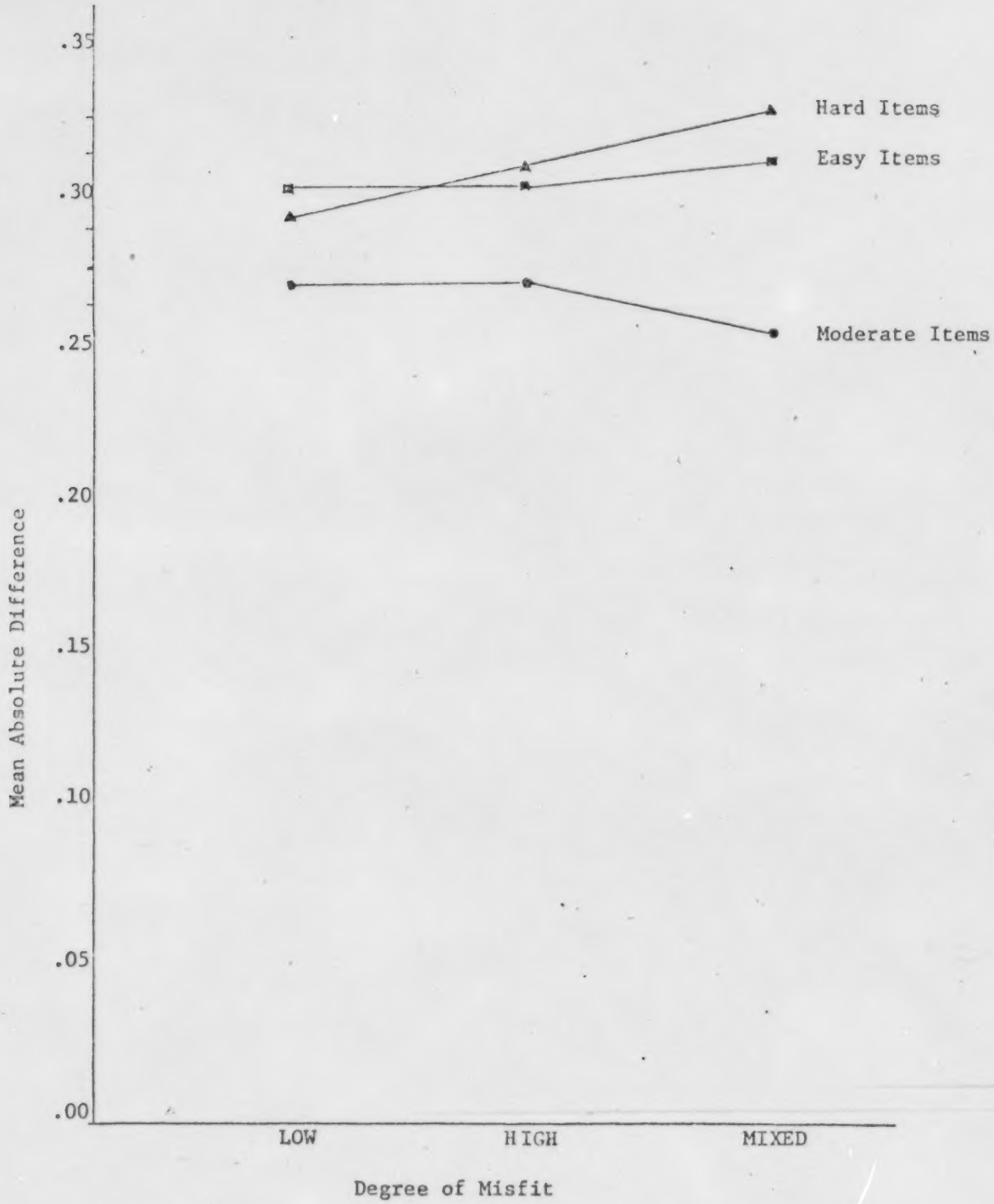

Figure 2. Cell Means for Interaction of Degree of Misfit and Difficulty

means, shows an increase in $|\hat{\theta} - \theta|$ for moderate items. The expectation would be that, as the percentage of items conforming to the Rasch model increases, the correspondence between ability estimates and true abilities would increase (lower values of $|\hat{\theta} - \theta|$).

Explanations for these unexpected results were sought with the following areas investigated: errors in item subset labeling or construction, data simulation program errors, and misuse or improper calculation of the dependent variable, $|\hat{\theta} - \theta|$. It was found that the item subsets were correctly constructed and labeled as outlined. A detailed review of the simulation program found no programming, logical, or computational errors. Furthermore, the basic-simulation program has been successfully employed in at least two other Rasch model investigations (Cartledge, 1974; C. C. Rentz, 1975). Consideration of the dependent variable yielded a similar outcome: the variable was calculated and used properly, and it has been successfully employed in at least one previous study (C. C. Rentz, 1975). Thus the readily apparent sources of error were accounted for but with no solution to the problem.

Next, a basic review of the measurement process was undertaken in order to identify other, less obvious, sources of error. This review was based upon the question: what is the expectation when we obtain an ability estimate vis-a-vis the ability parameter? In response, the viewpoint taken was that, when we obtain an ability estimate, we do not expect exact numerical correspondence between the parameter and this estimate. However, if the measurement is adequate, we would expect a confidence interval based upon error of measurement and the ability estimate to "capture" the parameter. In the event that the measurement-error interval around the estimate includes the para-

meter, we would conclude parameter and estimate correspondence and, thus, also conclude invariance.

As shown previously, the minimum standard error of measurement may be constructed if all item difficulties match the ability estimate (i.e., $\hat{\theta}_j = d_i$, for all i). Thus the standard error formulation, equation (1), reduces to

$$SEM_j(\min) = \sqrt{1/L(.25)}$$

where L is the number of items in the test or subset. As employed in this study, a subset length of 40 items produces a minimum SEM of approximately .32. The corresponding 68% confidence interval would be $\hat{\theta} \pm .32$. That is, we can expect $\hat{\theta}$ to "fluctuate" at least $\pm$ .32 log units around θ, a fluctuation due to measurement error alone. However, when using absolute differences, this fluctuation is reduced to positive values only. Thus, values of $|\hat{\theta} - \theta|$ that are .32 or less would indicate correspondence within the bounds of minimum measurement error. It should be noted that these values are quite conservative; when the difficulty-ability match deviates from optimum ($\hat{\theta}_j \neq d_i$, for some, or all i) the standard error of the estimate will increase.

Employing .32 as a rough indicator of the upper bound of invariance, the data results (Tables 4, 5, and 6) may be reinterpreted. The largest mean absolute difference is .35, and only 17% of the 108 ability-estimation situations produced mean absolute differences greater than .32. The data indicate that, using a very stringent criterion, the Rasch model yields ability estimates that closely correspond to true abilities. Furthermore, a viewpoint that may be adopted is that the simulation results shows only random fluctuations within a band that indicates invariance.

Inferring from the above interpretation, we would expect that situations similar to those studied would yield invariant Rasch-model ability estimates. If this contention were true, then alteration of the ability estimation situation might produce a lack of invariance. A key to such an alteration is found in the Hambleton and Cook (1977) representation of the Rasch model. The Hambleton and Cook (1977, p. 12) presentation of the Rasch model can be shown to be equivalent to

$$P_{ij} = \frac{e^{D\bar{a}(\theta_j - d_i)}}{1 + e^{D\bar{a}(\theta_j - d_i)}} \qquad (5)$$

where $\bar{a}$ is the average item discrimination over the items in the test and D is the scaling factor of 1.7 (see Appendix A). The D is of little interest in that it serves merely as a factor which brings correspondence between logistic models and the normal-ogive model. However, the use of $\bar{a}$ is more intriguing. As written, equation (5) indicates that the average of the test-item discriminations serves as a "scaling" factor for model parameter estimates. This suggests that, if we alter $\bar{a}$ for the subsets of this investigation, then the resulting ability estimates may not conform to the ability parameters. However, this lack of conformity should be due only to scale differences between the ability estimates and true abilities. Therefore, if we rescale the estimate by $\bar{a}$, these adjusted estimates should conform to the true abilities.

Additional Simulations

To ascertain the effect of altering average item discrimination, two additional simulations were performed. In order to avoid negative values for some item discriminations, $\bar{a}$ was increased rather than

decreased. The value of the average discrimination was set to 1.5; however, discriminations of these magnitudes could result in many examinees passing or failing all items in a subset. To circumvent this problem, the true-ability interval was expanded to approximately $\pm$ 3.33. The average item discrimination was altered by adding .5 to the original discriminations.

The first of the two additional simulations employed the expanded interval of true ability but maintained all other simulation factors (e.g., $\bar{a} = 1.0$). The purpose of this simulation was to determine baseline data under the broader range of true ability. The simulation results summary, Tables 11, 12, and 13 show that all but one mean absolute difference to be above .32 for the normal distribution group. In contrast 64% of these means are above .32 for the positive and negative skew groups of examinees. The ranges of mean absolute differences for the normal, positive skew, and negative skew examinee groups were .306 to .604, .247 to .387, and .258 to .400, respectively.

It is quite clear that the simulation results for the expanded range of ability are less satisfactory than the results of the original simulation. In explanation it should be noted that for the original ability interval ($-1.0 \leq \theta \leq 1.0$) an average item subset discrimination near unity is quite satisfactory. However, when we are trying to measure across nearly seven log units of ability, an average discrimination of unity is entirely too low to expect a high precision of measurement. Thus, if we desire the same measurement precision for a broad range of ability as that achieved for a narrow ability range, we must increase the discrimination level of our test. The above highlights Panchapakesan's statement that, under the Rasch model, the contention

is not that item quality (discrimination) is unimportant, but rather we desire items to be of similar quality (1969, p. 46).

These results also serve to point out two additional factors. The first is that the minimum standard error, equation (2) is not a totally satisfactory index for summarizing measurement precision when large ranges of ability are involved. The minimum standard error would be a useful summary if we are employing items appropriate in difficulty and if the item difficulty interval closely matches the interval of abilities to be measured. In situations other than this the assumption that all P_{ij} are .5 becomes totally unsupported by the data and the minimum standard error provides an overly conservative estimate of the average standard error of measurement.

The second factor is that the expansion of the ability range serves to accentuate the differences between the levels-of-item subset difficulty studied. In almost all instances, when the percentage of fitting items in the subset and degree of subset misfit are held constant, the ordering shows moderate items to be preferable to easy or hard items. These results may be attributed to moderate difficulty items "covering" more examinees with more items. That is, since the moderate items are centered on the average ability of the three examinee groups, we are able to obtain more precise measurement.

As noted earlier, these results do not appear as satisfactory as the original results. Table 14 serves to note that we are obtaining adequate orderings of our examinees as to ability; however, this same table also serves to point out the limitations of the correlation coefficient in invariance studies. The normal-distribution group shows high correlations for ability estimates and true abilities, but the absolute difference of these two variables is quite high. The

Table 11

Simulation Results for Normal Distribution

of True Ability (Expanded Ability Interval)

True Ability (θ)

$\bar{X}$ = -0.220 sd = 1.781 min = -3.220 max = 3.170

Item Subset* P, M, D	Observed Score $\bar{X}$	sd	min	max	Rasch Model $\hat{\theta}$ $\bar{X}$	sd	min	max	$\lvert\hat{\theta} - \theta\rvert$ $\bar{X}$	sd	min	max
0%, Low, Easy	24.11	11.514	0	39	- .386	1.685	-4.763	2.763	.497	.382	.005	1.718
0%, Low, Moderate	18.73	11.841	0	39	- .187	1.720	-3.763	3.763	.306	.250	.003	1.046
0%, Low, Hard	13.05	10.691	0	36	- .100	1.640	-2.763	3.281	.382	.369	.007	1.614
0%, High, Easy	23.86	10.252	5	38	- .479	1.323	-3.025	2.038	.508	.513	.005	2.415
0%, High, Moderate	19.21	10.445	1	37	- .124	1.374	-3.763	2.601	.479	.397	.002	1.528
0%, High, Hard	14.09	9.929	2	35	.168	1.343	-2.039	3.024	.604	.549	.000	2.355
0%, Mixed, Easy	23.77	10.800	3	38	- .475	1.418	-3.601	2.038	.468	.453	.008	1.952
0%, Mixed, Moderate	18.73	11.389	2	38	- .207	1.538	-3.039	3.038	.396	.351	.003	1.503
0%, Mixed, Hard	13.55	10.392	0	35	.017	1.510	-2.763	3.024	.515	.442	.001	1.789
25%, Low, Easy	24.03	12.033	1	40	- .406	1.750	-4.763	2.763	.409	.365	.011	1.962
25%, Low, Moderate	18.84	12.438	0	39	- .193	1.799	-3.763	3.763	.380	.273	.011	1.166
25%, Low, Hard	13.16	10.429	0	37	- .051	1.564	-2.763	3.601	.368	.312	.023	1.259
25%, High, Easy	24.05	10.909	3	40	- .455	1.482	-3.601	2.763	.528	.451	.001	2.136
25%, High, Moderate	18.96	10.969	2	38	- .140	1.469	-3.039	3.038	.435	.329	.008	1.362
25%, High, Hard	14.00	10.320	0	38	.116	1.519	-2.763	4.038	.509	.489	.009	2.195
25%, Mixed, Easy	24.13	11.154	3	39	- .364	1.575	-3.601	2.763	.435	.408	.009	2.019
25%, Mixed, Moderate	19.07	11.265	1	38	- .164	1.550	-3.763	3.038	.399	.314	.008	1.412
25%, Mixed, Hard	13.15	10.164	0	35	- .039	1.514	-2.763	3.024	.452	.444	.002	2.166
50%, Low, Easy	24.69	11.479	3	40	- .250	1.659	-3.601	2.763	.341	.269	.002	1.353
50%, Low, Moderate	18.61	12.225	1	38	- .279	1.783	-3.763	3.038	.341	.297	.006	1.542
50%, Low, Hard	12.92	10.921	0	37	- .167	1.724	-2.763	3.601	.385	.335	.000	1.534
50%, High, Easy	24.40	11.305	3	39	- .360	1.591	-3.601	2.763	.399	.346	.002	1.916
50%, High, Moderate	18.85	11.592	2	39	- .172	1.604	-3.039	3.763	.381	.312	.003	1.466
50%, High, Hard	13.55	10.457	0	36	.007	1.566	-2.763	3.281	.496	.429	.017	2.102
50%, Mixed, Easy	24.35	11.374	3	40	- .358	1.564	-3.601	2.763	.415	.401	.002	1.763
50%, Mixed, Moderate	18.41	11.839	1	40	- .246	1.725	-3.763	3.763	.374	.297	.000	1.372
50%, Mixed, Hard	13.49	10.718	0	36	- .042	1.643	-2.763	3.281	.404	.349	.047	1.696

Table 11 continued

Item Subset* P, H, D	Observed Score $\bar{X}$	sd	min	max	Rasch Model $\hat{\theta}$ $\bar{X}$	sd	min	max	$\lvert\hat{\theta} - \theta\rvert$ $\bar{X}$	sd	min	max
75%, Low, Easy	25.03	11.958	1	40	- .165	1.842	-4.763	2.763	.409	.404	.001	2.115
75%, Low, Moderate	18.92	12.240	1	40	- .174	1.761	-3.763	3.763	.403	.313	.032	1.595
75%, Low, Hard	12.98	11.043	0	26	- .137	1.704	-2.763	3.281	.389	.336	.009	1.423
75%, High, Easy	24.29	11.001	3	39	- .339	1.570	-3.601	2.763	.386	.316	.005	1.432
75%, High, Moderate	18.34	11.893	0	40	- .261	1.740	-3.763	3.763	.362	.293	.000	1.475
75%, High, Hard	13.32	10.872	0	37	- .020	1.583	-2.763	3.601	.407	.376	.001	1.776
75%, Mixed, Easy	25.02	11.435	3	40	- .231	1.676	-3.601	2.763	.348	.283	.002	1.210
75%, Mixed, Moderate	18.69	12.116	0	38	- .231	1.746	-3.763	3.038	.366	.291	.014	1.229
75%, Mixed, Hard	12.89	10.355	0	35	- .106	1.586	-2.763	3.024	.355	.330	.000	1.789

*P = Percentage of Fitting Items, M = Degree of Misfit, D = Difficulty Level

Table 12

Simulation Results for Positively Skewed Distributi

of True Ability (Expanded Ability Interval)

True Ability (θ)

$\bar{X}$ = -.057 sd = 0.964 min = -.999 max = 2.687

Item Subset P, M, D	Observed Score $\bar{X}$	sd	min	max	Rasch Model $\hat{\theta}$ $\bar{X}$	sd	min	max	$\lvert\hat{\theta} - \theta\rvert$ $\bar{X}$	sd	min	max
0%, Low, Easy	27.76	6.220	14	40	-.115	.945	-1.651	2.763	.350	.396	.004	1.258
0%, Low, Moderate	19.30	7.507	8	37	-.047	.919	-1.449	2.601	.260	.223	.001	1.091
0%, Low, Hard	12.01	7.724	0	35	-.026	1.039	-2.763	3.024	.291	.282	.001	1.764
0%, High, Easy	25.85	5.918	14	38	-.283	.771	-1.651	2.038	.386	.348	.012	1.655
0%, High, Moderate	19.24	7.399	9	36	-.065	.868	-1.295	2.281	.290	.218	.018	1.055
0%, High, Hard	12.76	7.268	3	32	.109	.911	-1.601	2.449	.363	.272	.001	1.075
0%, Mixed, Easy	26.59	6.229	15	39	-.151	.899	-1.537	2.736	.350	.312	.000	1.696
0%, Mixed, Moderate	18.93	8.088	5	37	-.093	.992	-2.025	2.601	.276	.245	.002	1.091
0%, Mixed, Hard	12.38	7.594	1	34	.023	1.037	-2.763	2.808	.387	.336	.007	2.024
25%, Low, Easy	26.44	6.495	15	39	-.156	.939	-1.537	2.763	.352	.275	.001	1.696
25%, Low, Moderate	18.61	8.329	7	39	-.087	1.142	-1.619	3.763	.318	.264	.007	1.758
25%, Low, Hard	12.24	7.273	3	32	.027	.934	-1.601	2.449	.296	.216	.006	.962
25%, High, Easy	26.69	5.346	17	39	-.174	.768	-1.318	2.763	.369	.384	.000	1.922
25%, High, Moderate	19.19	7.712	8	35	-.067	.918	-1.449	2.024	.267	.201	.001	1.044
25%, High, Hard	12.77	7.894	2	34	.093	.992	-2.039	2.808	.344	.245	.001	1.319
25%, Mixed, Easy	26.83	6.357	14	38	-.126	.886	-1.651	2.038	.376	.350	.004	1.860
25%, Mixed, Moderate	19.06	7.769	6	38	-.072	.968	-1.808	3.038	.247	.197	.001	.952
25%, Mixed, Hard	12.27	7.452	2	34	.015	.979	-2.039	2.808	.334	.238	.012	1.195
50%, Low, Easy	27.10	6.428	15	40	-.055	.994	-1.537	2.763	.291	.200	.011	1.141
50%, Low, Moderate	19.10	8.205	7	38	-.063	1.029	-1.619	3.038	.269	.201	.003	.996
50%, Low, Hard	11.78	7.262	2	31	-.054	.961	-2.039	2.295	.328	.266	.001	1.183
50%, High, Easy	26.71	6.117	13	39	-.054	.904	-1.768	2.763	.331	.259	.000	1.240
50%, High, Moderate	19.11	7.318	8	37	-.083	.859	-1.449	2.601	.306	.271	.001	1.246
50%, High, Hard	12.17	7.665	3	35	.014	.991	-1.601	3.024	.329	.250	.002	1.346
50%, Mixed, Easy	26.77	6.356	14	40	-.092	1.011	-1.651	2.763	.343	.272	.004	1.191
50%, Mixed, Moderate	18.82	8.095	7	38	-.097	1.005	-1.619	3.039	.383	.177	.003	.844
50%, Mixed, Hard	12.02	7.64	2	35	-.023	1.044	-2.039	3.024	.387	.269	.004	1.050

Table 12 continued

Item Subset* P, H, D	Observed Score $\bar{X}$	sd	min	max	Rasch Model $\hat{\theta}$ $\bar{X}$	sd	min	max	$\lvert\hat{\theta} - \theta\rvert$ $\bar{X}$	sd	min	max
75%, Low, Easy	26.93	6.717	14	40	-.056	1.053	-1.651	2.763	.332	.296	.003	1.346
75%, Low, Moderate	19.16	8.069	5	39	-.064	1.011	-2.025	3.763	.334	.252	.010	1.222
75%, Low, Hard	11.76	8.244	2	35	-.094	1.133	-2.039	3.024	.364	.284	.006	1.249
75%, High, Easy	26.78	5.904	16	40	-.132	.873	-1.427	2.763	.343	.271	.003	1.266
75%, High, Moderate	19.11	7.870	7	38	-.058	.993	-1.619	3.038	.274	.207	.007	1.055
75%, High, Hard	11.45	7.542	1	33	-.127	1.058	-2.763	2.619	.360	.331	.003	1.777
75%, Mixed, Easy	27.46	6.497	16	40	-.002	1.002	-1.427	2.763	.316	.239	.011	1.191
75%, Mixed, Moderate	19.14	8.088	5	38	-.065	1.004	-2.025	3.038	.294	.225	.001	1.169
75%, Mixed, Hard	11.69	8.140	2	37	-.083	1.102	-2.039	3.601	.324	.269	.007	1.249

*P = Percentage of Fitting Items, H = Degree of Misfit, D = Difficulty Level

Table 13

Simulation Results for Negatively Skewed Distribution of True Ability (Expanded Ability Interval)

True Ability (θ)

$\bar{X}$ = .131 sd = .972 min = -2.781 max = .999

Item Subset* P, H, D	Observed Score $\bar{X}$	sd	min	max	Rasch Model $\hat{\theta}$ $\bar{X}$	sd	min	max	$\lvert\hat{\theta} - \theta\rvert$ $\bar{X}$	sd	min	max
0%, Low, Easy	28.55	7.663	6	39	.103	1.018	-2.808	2.763	.370	.313	.008	1.780
0%, Low, Moderate	21.71	8.228	0	33	.128	1.082	-3.763	1.619	.300	.234	.005	1.455
0%, Low, Hard	13.44	6.027	0	29	.153	.907	-2.763	2.017	.289	.259	.002	1.044
0%, High, Easy	27.74	7.192	5	39	-.045	.931	-3.025	2.763	.394	.317	.001	2.140
0%, High, Moderate	21.75	7.224	5	33	.171	.866	-2.025	1.719	.297	.217	.000	.972
0%, High, Hard	14.58	6.015	3	25	.316	.824	-1.601	1.537	.346	.371	.000	1.839
0%, Mixed, Easy	28.56	7.327	6	38	.078	.958	-2.808	2.038	.361	.271	.000	1.209
0%, Mixed, Moderate	21.42	7.620	2	35	.131	.943	-3.039	2.024	.275	.243	.002	1.055
0%, Mixed, Hard	14.46	6.433	1	25	.262	.967	-2.763	1.537	.400	.372	.003	1.756
25%, Low, Easy	28.11	7.760	5	38	.028	1.019	-3.025	2.038	.332	.227	.006	1.066
25%, Low, Moderate	21.24	7.993	2	31	.082	1.030	-3.039	1.295	.304	.241	.005	1.420
25%, Low, Hard	13.94	6.136	1	25	.217	.895	-2.763	1.537	.314	.225	.006	.993
25%, High, Easy	27.63	6.986	7	36	-.062	.875	-2.619	1.281	.400	.262	.004	1.198
25%, High, Moderate	21.70	7.724	4	34	.175	.931	-2.281	1.808	.307	.277	.005	1.576
25%, High, Hard	14.35	6.081	1	28	.287	.848	-2.763	1.839	.379	.334	.004	1.756
25%, Mixed, Easy	28.52	7.889	5	40	.103	1.056	-3.025	2.763	.348	.308	.006	1.790
25%, Mixed, Moderate	21.83	7.992	3	34	.178	.983	-2.601	1.808	.258	.186	.000	.833
25%, Mixed, Hard	14.38	6.405	1	26	.258	.949	-2.763	1.651	.358	.282	.004	1.179
50%, Low, Easy	28.82	8.121	6	39	.168	1.105	-2.808	2.763	.314	.307	.002	1.760
50%, Low, Moderate	21.95	8.407	0	33	.169	1.099	-3.763	1.719	.291	.264	.007	1.788
50%, Low, Hard	13.38	5.984	0	25	.126	.956	-2.763	1.537	.297	.267	.000	1.435
50%, High, Easy	28.50	7.694	7	37	.088	.997	-2.619	1.601	.317	.237	.001	.949
50%, High, Moderate	21.81	7.585	1	33	.154	.983	-3.763	1.619	.277	.225	.010	.982
50%, High, Hard	14.17	6.169	1	26	.236	.938	-2.763	1.651	.310	.276	.004	1.433
50%, Mixed, Easy	28.42	7.560	5	37	.067	.996	-3.025	1.701	.345	.249	.000	1.050
50%, Mixed, Moderate	21.06	7.875	1	33	.067	1.030	-3.763	1.619	.328	.230	.002	.982
50%, Mixed, Hard	13.67	6.568	1	27	.123	1.072	-2.763	1.768	.317	.251	.004	1.158

Table 13 continued

Item Subset* P, H, D	Observed Score $\bar{X}$	sd	min	max	Rasch Model $\hat{\theta}$ $\bar{X}$	sd	min	max	$\lvert\hat{\theta} - \theta\rvert$ $\bar{X}$	sd	min	max
75%, Low, Easy	28.95	7.527	6	39	.172	1.034	-2.808	2.763	.340	.342	.015	2.193
75%, Low, Moderate	21.47	8.110	2	35	.132	1.015	-3.039	2.024	.297	.230	.002	1.025
75%, Low, Hard	13.59	6.415	0	26	.123	1.045	-2.763	1.651	.318	.284	.001	1.333
75%, High, Easy	28.44	7.421	7	39	.077	.992	-2.619	2.763	.338	.322	.006	2.070
75%, High, Moderate	21.69	7.668	4	36	.169	.944	-2.281	2.281	.325	.266	.002	1.588
75%, High, Hard	13.54	6.267	0	26	.142	.973	-2.763	1.651	.303	.298	.004	1.839
75%, Mixed, Easy	28.77	7.486	5	38	.122	1.002	-3.025	2.038	.325	.275	.002	1.263
75%, Mixed, Moderate	21.55	8.139	3	35	.130	1.034	-2.601	2.024	.273	.237	.002	1.442
75%, Mixed, Hard	13.68	6.337	0	24	.162	.964	-2.763	1.426	.295	.228	.001	1.126

*P = Percentage of Fitting Items, H = Degree of Misfit, D = Difficulty Level

Table 14

Correlations and Regression Line Slopes for Expanded Interval of True Ability and Rasch Model Ability Estimates

Item Subset* P, M, D	True Ability Distribution Shape					
	Normal Distribution		Positive Skew		Negative Skew	
	r	Slope	r	Slope	r	Slope
0%, Low, Easy	.94	1.00	.89	0.90	.88	0.84
0%, Low, Moderate	.98	1.01	.93	0.98	.94	0.84
0%, Low, Hard	.96	1.04	.92	0.85	.92	0.98
0%, High, Easy	.95	1.28	.88	1.10	.88	0.91
0%, High, Moderate	.96	1.24	.93	1.03	.93	1.04
0%, High, Hard	.93	1.24	.90	0.95	.87	1.03
0%, Mixed, Easy	.95	1.20	.88	0.94	.89	0.90
0%, Mixed, Moderate	.96	1.11	.93	0.90	.93	0.96
0%, Mixed, Hard	.94	1.11	.87	0.81	.85	0.85
25%, Low, Easy	.96	0.97	.89	0.92	.92	0.88
25%, Low, Moderate	.97	0.96	.94	0.79	.93	0.87
25%, Low, Hard	.97	1.11	.93	0.96	.92	1.00
25%, High, Easy	.94	1.12	.84	1.06	.89	0.99
25%, High, Moderate	.96	1.17	.94	0.98	.91	0.95
25%, High, Hard	.94	1.10	.92	0.89	.87	1.00
25%, Mixed, Easy	.95	1.07	.85	0.93	.90	0.83
25%, Mixed, Moderate	.96	1.11	.95	0.94	.95	0.94
25%, Mixed, Hard	.94	1.11	.91	0.90	.90	0.92
50%, Low, Easy	.97	1.04	.94	0.91	.92	0.81
50%, Low, Moderate	.97	0.97	.94	0.86	.94	0.83
50%, Low, Hard	.96	0.99	.90	0.91	.91	0.83
50%, High, Easy	.96	1.08	.90	0.96	.92	0.90
50%, High, Moderate	.96	1.07	.91	1.02	.93	0.92
50%, High, Hard	.94	1.07	.91	0.89	.91	0.94
50%, Mixed, Easy	.95	1.08	.90	0.86	.91	0.87
50%, Mixed, Moderate	.96	0.99	.94	0.91	.92	0.87
50%, Mixed, Hard	.96	1.04	.89	0.82	.93	0.84
75%, Low, Easy	.95	0.92	.91	0.83	.89	0.84
75%, Low, Moderate	.96	0.97	.91	0.87	.93	0.89
75%, Low, Hard	.96	1.00	.92	0.78	.91	0.85
75%, High, Easy	.97	1.10	.89	0.99	.89	0.87
75%, High, Moderate	.97	0.99	.94	0.91	.90	0.93
75%, High, Hard	.96	1.08	.89	0.81	.90	0.90
75%, Mixed, Easy	.97	1.03	.92	0.89	.91	0.88
75%, Mixed, Moderate	.96	0.98	.93	0.89	.94	0.88
75%, Mixed, Hard	.97	1.09	.93	0.81	.93	0.93

*P = Percentage of Fitting Items, M = Degree of Misfit, D = Difficulty Level

opposite trend is found for the other two examinee groups. These results stem from the correlation coefficient's sensitivity to range restrictions in the correlated variables. A restricted range serves to lower the correlation, whereas the same restriction serves to improve our measurement as reflected in the absolute difference data.

With the above as background, the second supplementary simulation was performed. These results, as summarized in Tables 15, 16, and 17, were obtained using the expanded range of ability and an average item discrimination of 1.5. In each table the columns of greatest interest are the unadjusted and adjusted absolute differences. In general the unadjusted data indicate that the correspondence between estimates and parameters is lessened considerably (i.e., higher $|\hat{\theta} - \theta|$ values). However, when $\hat{\theta}$ is rescaled by an estimate of the average item discrimination (adjusted $\hat{\theta} = \hat{\theta}/1.5$), the absolute-difference data indicates greater correspondence.

A comparison of these adjusted absolute differences and the absolute differences obtained from using $\bar{a}$ of 1.0 indicates that the use of $\bar{a}$ as a scaling factor is quite helpful.

Table 15

Simulation Results Summary For Normal Distribution of True True Ability

(Expanded Interval of True Ability and $\bar{a}$ = 1.5)

True Ability (θ)

$\bar{X}$ = .220 sd = 1.701

Item Subset* P,M,D	Observed Score $\bar{X}$	Observed Score sd	Rasch Model $\hat{\theta}$ $\bar{X}$	Rasch Model $\hat{\theta}$ sd	$\hat{\theta}/1.5$ $\bar{X}$	$\hat{\theta}/1.5$ sd	Unadjusted $\lvert\hat{\theta} - \theta\rvert$ $\bar{X}$	Unadjusted $\lvert\hat{\theta} - \theta\rvert$ sd	Unadjusted $\lvert(\hat{\theta}/1.5)-\theta\rvert$ $\bar{X}$	Unadjusted $\lvert(\hat{\theta}/1.5)-\theta\rvert$ sd
0%, Low, Easy	23.88	13.658	-.135	2.236	-.090	1.491	.528	.485	.439	.329
0%, Low, Moderate	18.51	14.306	-.271	2.350	-.181	1.567	.544	.392	.282	.229
0%, Low, Hard	12.25	12.633	-.365	2.151	-.244	1.434	.521	.434	.445	.353
0%, High, Easy	24.89	13.182	-.160	2.063	-.106	1.375	.410	.331	.441	.335
0%, High, Moderate	18.78	13.148	-.253	2.051	-.168	1.367	.397	.311	.420	.342
0%, High, Hard	12.57	12.187	-.328	2.001	-.219	1.334	.427	.364	.477	.376
0%, Mixed, Easy	25.13	13.792	-.089	2.238	-.059	1.492	.565	.473	.464	.326
0%, Mixed, Moderate	18.54	14.123	-.327	2.301	-.218	1.534	.555	.436	.342	.298
0%, Mixed, Hard	12.46	12.434	-.342	2.056	-.228	1.371	.473	.431	.486	.379
25%, Low, Easy	24.75	14.311	-.194	2.339	-.130	1.559	.631	.438	.396	.296
25%, Low, Moderate	18.87	14.563	-.206	2.437	-.137	1.624	.619	.448	.300	.213
25%, Low, Hard	12.11	12.584	-.359	2.131	-.239	1.421	.506	.420	.459	.335
25%, High, Easy	24.90	13.413	-.197	2.110	-.131	1.407	.481	.430	.472	.359
25%, High, Moderate	18.80	13.851	-.218	2.225	-.145	1.483	.467	.381	.348	.290
25%, High, Hard	12.45	12.260	-.303	2.058	-.202	1.372	.417	.409	.461	.355
25%, Mixed, Easy	24.92	13.689	-.112	2.253	-.075	1.502	.545	.462	.419	.330
25%, Mixed, Moderate	18.76	13.999	-.278	2.280	-.185	1.520	.509	.397	.329	.267
25%, Mixed, Hard	11.73	12.262	-.463	2.060	-.309	1.373	.538	.497	.551	.368
50%, Low, Easy	25.21	13.921	-.030	2.281	-.020	1.521	.586	.517	.459	.338
50%, Low, Moderate	18.60	14.450	-.313	2.395	-.209	1.597	.605	.411	.279	.240
50%, Low, Hard	11.95	12.691	-.453	2.183	-.302	1.456	.568	.488	.478	.337
50%, High, Easy	25.32	13.657	-.077	2.213	-.051	1.476	.523	.448	.462	.294
50%, High, Moderate	18.62	14.288	-.287	2.368	-.191	1.579	.579	.439	.310	.260
50%, High, Hard	12.23	12.490	-.396	2.101	-.264	1.400	.523	.418	.475	.360
50%, Mixed, Easy	25.09	14.045	-.096	2.256	-.064	1.504	.567	.451	.438	.314
50%, Mixed, Moderate	18.38	14.191	-.337	2.332	-.225	1.555	.551	.441	.327	.255
50%, Mixed, Hard	12.17	12.566	-.422	2.131	-.281	1.421	.527	.451	.479	.341
75%, Low, Easy	25.55	14.135	.008	2.333	.006	1.556	.658	.527	.495	.304
75%, Low, Moderate	18.62	14.457	-.296	2.359	-.198	1.573	.598	.446	.332	.283
75%, Low, Hard	11.99	12.408	-.442	2.109	-.295	1.406	.526	.485	.508	.352
75%, High, Easy	25.02	13.531	-.107	2.227	-.071	1.484	.516	.470	.446	.307
75%, High, Moderate	18.43	14.401	-.303	2.388	-.202	1.592	.597	.455	.304	.265
75%, High, Hard	12.10	12.766	-.404	2.122	-.269	1.415	.539	.401	.463	.330
75%, Mixed, Easy	25.64	13.481	-.041	2.239	-.027	1.492	.549	.494	.477	.309
75%, Mixed, Moderate	18.55	14.361	-.258	2.400	-.172	1.600	.600	.477	.309	.275
75%, Mixed, Hard	11.87	12.323	-.436	2.102	-.290	1.401	.524	.412	.471	.351

* P = Percentage of Fitting Items, M = Degree of Misfit, D = Difficulty Level

Table 16

Simulation Results Summary for Positively Skewed

Distribution of True Ability

(Expanded Interval of True Ability and $\bar{a} = 1.5$)

True Ability (θ)

$\bar{X} = -.057$ sd = .964

Item Subset* P, M, D	Observed Score $\bar{X}$	Observed Score sd	Rasch Model $\hat{\theta}$ $\bar{X}$	Rasch Model $\hat{\theta}$ sd	$\hat{\theta}/1.5$ $\bar{X}$	$\hat{\theta}/1.5$ sd	Unadjusted $\lvert\theta - \hat{\theta}\rvert$ $\bar{X}$	Unadjusted $\lvert\theta - \hat{\theta}\rvert$ sd	$\lvert(\hat{\theta}/1.5)-\theta\rvert$ $\bar{X}$	$\lvert(\hat{\theta}/1.5)-\theta\rvert$ sd
0%, Low, Easy	28.58	7.145	.259	1.267	.179	.845	.440	.425	.359	.242
0%, Low, Moderate	18.90	10.005	-.032	1.398	-.021	.932	.403	.324	.185	.147
0%, Low, Hard	9.34	9.430	-.539	1.481	-.359	.988	.636	.537	.370	.290
0%, High, Easy	27.88	7.151	.114	1.150	.076	.767	.341	.319	.320	.239
0%, High, Moderate	18.66	9.596	-.109	1.240	-.073	.827	.361	.241	.230	.201
0%, High, Hard	10.04	9.301	-.452	1.384	-.301	.923	.542	.461	.350	.234
0%, Mixed, Easy	28.44	7.283	.228	1.224	.152	.816	.406	.391	.345	.242
0%, Mixed, Moderate	18.34	10.067	-.132	1.325	-.088	.883	.386	.304	.207	.190
0%, Mixed, Hard	9.96	9.386	-.462	1.466	-.308	.977	.599	.495	.341	.271
25%, Low, Easy	28.26	7.389	.249	1.314	.166	.876	.461	.300	.361	.303
25%, Low, Moderate	18.21	10.477	-.111	1.499	-.074	.999	.467	.401	.190	.163
25%, Low, Hard	10.00	8.998	-.437	1.328	-.291	.885	.486	.397	.296	.228
25%, High, Easy	28.59	6.344	.204	1.117	.136	.745	.350	.294	.339	.221
25%, High, Moderate	18.46	9.885	-.119	1.316	-.079	.877	.362	.294	.205	.163
25%, High, Hard	10.44	9.544	-.405	1.405	-.270	.937	.552	.421	.306	.253
25%, Mixed, Easy	28.55	7.174	.253	1.236	.168	.824	.452	.440	.382	.277
25%, Mixed, Moderate	18.29	10.199	-.122	1.420	-.081	.946	.422	.333	.186	.137
25%, Mixed, Hard	10.01	9.195	-.517	1.479	-.345	.986	.633	.563	.414	.271
50%, Low, Easy	28.91	7.355	.336	1.304	.224	.870	.516	.418	.379	.241
50%, Low, Moderate	18.42	10.577	-.109	1.477	-.073	.984	.432	.388	.196	.150
50%, Low, Hard	9.40	9.241	-.625	1.460	-.417	.973	.679	.553	.409	.309
50%, High, Easy	28.59	7.031	.236	1.200	.158	.800	.405	.396	.371	.232
50%, High, Moderate	18.57	9.858	-.121	1.269	-.081	.846	.382	.281	.229	.220
50%, High, Hard	9.85	9.467	-.519	1.440	-.346	.960	.612	.516	.360	.297
50%, Mixed, Easy	28.65	7.102	.259	1.242	.173	.828	.444	.383	.357	.234
50%, Mixed, Moderate	18.43	10.400	-.090	1.451	-.060	.967	.455	.392	.198	.154
50%, Mixed, Hard	9.79	9.475	-.516	1.515	-.344	1.010	.634	.524	.380	.257
75%, Low, Easy	28.80	7.311	.314	1.307	.209	.871	.467	.471	.373	.252
75%, Low, Moderate	18.39	10.206	-.124	1.396	-.083	.931	.452	.359	.239	.200
75%, Low, Hard	9.86	9.777	-.569	1.353	-.379	1.035	.638	.568	.392	.283
75%, High, Easy	28.76	6.867	.274	..322	.183	.814	.441	.412	.372	.259
75%, High, Moderate	18.47	10.003	-.102	1.367	-.068	.911	.387	.343	.199	.180
75%, High, Hard	9.51	9.431	-.589	1.481	-.393	.981	.664	.487	.378	.251
75%, Mixed, Easy	28.88	7.217	.342	1.324	.228	.883	.524	.445	.393	.237
75%, Mixed, Moderate	18.54	10.381	-.096	1.419	-.064	.946	.420	.349	.181	.169
75%, Mixed, Hard	9.79	9.845	-.534	1.562	-.356	1.041	.681	.511	.356	.252

*P = Percentage of Fitting Items, M = Degree of Misfit, D = Difficulty Level

Table 17

Simulation Results Summary for Negatively Skewed

Distribution of True Ability

(Expanded Interval of True Ability and $\bar{x}$ = 1.5)

True Ability (θ)

$\bar{X}$ = .131 sd = 0.972

Item Subset* P, M, D	Observed Score $\bar{X}$	Observed Score sd	Rasch Model $\hat{\theta}$ $\bar{X}$	Rasch Model $\hat{\theta}$ sd	$\hat{\theta}/1.5$ $\bar{X}$	$\hat{\theta}/1.5$ sd	Unadjusted $\lvert\hat{\theta} - \theta\rvert$ $\bar{X}$	Unadjusted $\lvert\hat{\theta} - \theta\rvert$ sd	$\lvert(\hat{\theta}/1.5) - \theta\rvert$ $\bar{X}$	$\lvert(\hat{\theta}/1.5) - \theta\rvert$ sd
0%, Low, Easy	30.85	9.687	.652	1.474	.435	.982	.662	.516	.375	.271
0%, Low, Moderate	22.74	10.272	.196	1.436	.131	.991	.456	.375	.192	.132
0%, Low, Hard	11.98	6.863	-.135	1.169	-.090	.779	.394	.337	.372	.232
0%, High, Easy	30.44	9.288	.500	1.412	.334	.941	.527	.507	.337	.256
0%, High, Moderate	22.45	9.636	.189	1.356	.126	.904	.398	.365	.241	.201
0%, High, Hard	12.69	7.071	-.059	1.196	-.039	.797	.395	.364	.349	.269
0%, Mixed, Easy	30.72	9.162	.580	1.387	.386	.925	.587	.463	.356	.254
0%, Mixed, Moderate	22.15	9.987	.175	1.346	.117	.897	.372	.319	.196	.133
0%, Mixed, Hard	12.49	7.157	-.106	1.244	-.071	.829	.384	.476	.361	.306
25%, Low, Easy	30.49	9.360	.556	1.455	.371	.963	.595	.490	.333	.271
25%, Low, Moderate	22.38	10.145	.178	1.427	.119	.951	.435	.374	.227	.173
25%, Low, Hard	12.17	7.061	-.137	1.222	-.092	.815	.386	.333	.329	.223
25%, High, Easy	30.01	9.016	.419	1.239	.279	.860	.470	.412	.342	.236
25%, High, Moderate	22.46	10.092	.198	1.412	.132	.942	.424	.360	.212	.177
25%, High, Hard	12.50	7.150	-.079	1.208	-.053	.805	.415	.331	.347	.250
25%, Mixed, Easy	30.74	9.468	.617	1.473	.411	.982	.636	.528	.373	.295
25%, Mixed, Moderate	22.32	10.131	.198	1.421	.132	.947	.429	.339	.199	.154
25%, Mixed, Hard	12.64	7.342	-.075	1.241	-.050	.828	.394	.375	.334	.250
50%, Low, Easy	31.06	10.151	.751	1.641	.501	1.094	.818	.574	.410	.311
50%, Low, Moderate	22.76	10.278	.224	1.461	.149	.974	.446	.372	.185	.164
50%, Low, Hard	11.67	6.700	-.180	1.164	-.120	.776	.412	.349	.393	.223
50%, High, Easy	30.75	9.547	.622	1.476	.415	.984	.649	.479	.342	.260
50%, High, Moderate	22.34	9.616	.173	1.339	.116	.893	.363	.283	.185	.152
50%, High, Hard	12.30	7.291	-.091	1.226	-.060	.817	.381	.342	.330	.225
50%, Mixed, Easy	31.04	9.402	.674	1.499	.449	.999	.685	.520	.381	.263
50%, Mixed, Moderate	22.27	9.906	.193	1.355	.128	.904	.410	.312	.225	.163
50%, Mixed, Hard	12.17	7.037	-.155	1.260	-.103	.840	.404	.382	.346	.221
75%, Low, Easy	31.26	9.420	.764	1.508	.509	1.005	.745	.578	.431	.306
75%, Low, Moderate	22.34	10.190	.178	1.441	.119	.961	.432	.385	.213	.177
75%, Low, Hard	12.06	7.022	-.175	1.273	-.117	.849	.411	.431	.364	.239
75%, High, Easy	30.79	9.191	.584	1.392	.389	.928	.572	.480	.348	.256
75%, High, Moderate	22.53	9.806	.211	1.391	.140	.927	.406	.408	.234	.219
75%, High, Hard	11.86	6.969	-.174	1.211	-.116	.807	.404	.362	.364	.227
75%, Mixed, Easy	30.78	9.327	.702	1.507	.401	1.005	.626	.546	.344	.279
75%, Mixed, Moderate	22.36	9.962	.174	1.410	.116	.940	.392	.365	.204	.150
75%, Mixed, Hard	11.92	7.136	-.196	1.276	-.131	.851	.431	.400	.358	.229

*P = Percentage of Fitting Items, M = Degree of Misfit, D = Difficulty Level

Table 18

True Ability and Unadjusted and Adjusted Ability Estimate

Correlations and Regression Line Slopes

Item Subset*	Normal		θ	Positive Skew		θ	Negative Skew		θ
P, M, D	r	Unadjusted Slope	Adjusted Slope	r	Unadjusted Slope	Adjusted Slope	r	Unadjusted Slope	Adjusted Slope
0%, Low, Easy	.96	0.77	1.15	.93	0.71	1.06	.94	0.62	0.93
0%, Low, Moderate	.98	0.75	1.12	.97	0.67	1.00	.97	0.64	0.95
0%, Low, Hard	.96	0.80	1.19	.93	0.61	0.91	.93	0.77	1.16
0%, High, Easy	.97	0.84	1.26	.93	0.78	1.17	.92	0.64	0.95
0%, High, Moderate	.98	0.85	1.27	.95	0.74	1.11	.95	0.68	1.02
0%, High, Hard	.96	0.86	1.29	.93	0.65	0.98	.91	0.74	1.11
0%, Mixed, Easy	.96	0.76	1.15	.93	0.73	1.10	.93	0.65	0.98
0%, Mixed, Moderate	.97	0.75	1.13	.96	0.70	1.05	.96	0.69	1.04
0%, Mixed, Hard	.96	0.83	1.24	.93	0.61	0.92	.90	0.70	1.05
25%, Low, Easy	.97	0.74	1.10	.90	0.66	0.99	.93	0.63	0.94
25%, Low, Moderate	.98	0.72	1.08	.97	0.62	0.93	.96	0.65	0.98
25%, Low, Hard	.96	0.80	1.21	.95	0.69	1.04	.95	0.75	1.13
25%, High, Easy	.96	0.81	1.21	.94	0.82	1.22	.92	0.69	1.04
25%, High, Moderate	.98	0.78	1.13	.96	0.71	1.06	.96	0.66	0.99
25%, High, Hard	.96	0.83	1.25	.94	0.64	0.97	.92	0.74	1.11
25%, Mixed, Easy	.97	0.76	1.14	.90	0.71	1.06	.92	0.61	0.91
25%, Mixed, Moderate	.98	0.77	1.15	.97	0.66	0.99	.97	0.66	0.99
25%, Mixed, Hard	.95	0.82	1.23	.91	0.60	0.89	.92	0.72	1.09
50%, Low, Easy	.96	0.75	1.12	.93	0.69	1.03	.95	0.56	0.84
50%, Low, Moderate	.98	0.73	1.10	.97	0.63	0.95	.97	0.64	0.97
50%, Low, Hard	.96	0.78	1.17	.93	0.61	0.62	.93	0.78	1.17
50%, High, Easy	.97	0.78	1.17	.92	0.74	1.11	.95	0.62	0.93
50%, High, Moderate	.98	0.74	1.10	.95	0.72	1.08	.97	0.70	1.06
50%, High, Hard	.96	0.81	1.22	.93	0.62	0.93	.94	0.74	1.11
50%, Mixed, Easy	.96	0.76	1.14	.93	0.72	1.08	.94	0.61	0.92
50%, Mixed, Moderate	.98	0.75	1.12	.97	0.64	0.96	.96	0.69	1.03
50%, Mixed, Hard	.96	0.80	1.20	.93	0.59	0.89	.94	0.73	1.09
75%, Low, Easy	.96	0.73	1.10	.93	0.68	1.03	.93	0.60	0.90
75%, Low, Moderate	.97	0.74	1.10	.95	0.65	0.98	.96	0.65	0.97
75%, Low, Hard	.95	0.80	1.21	.94	0.58	0.87	.93	0.71	1.07
75%, High, Easy	.96	0.77	1.16	.92	0.73	1.09	.93	0.65	0.98
75%, High, Moderate	.98	0.73	1.09	.96	0.68	1.02	.94	0.66	0.99
75%, High, Hard	.96	0.81	1.21	.95	0.62	0.93	.94	0.75	1.13
75%, Mixed, Easy	.96	0.77	1.15	.93	0.68	1.01	.94	0.60	0.91
75%, Mixed, Moderate	.98	0.72	1.09	.97	0.66	0.99	.97	0.67	1.00
75%, Mixed, Hard	.96	0.81	1.22	.95	0.59	0.88	.94	0.82	1.07

* P = Perentage of Fitting Items, M = Degree of misfit, D = Difficulty Level

Utilizing Average Item Discrimination

In selecting a subset of items from a calibrated item pool, it is quite likely that the average item discrimination for this subset will differ from the total-item pool average. Thus ability estimates from this subset will not be on a scale that is equivalent to the ability estimate scale of the total-item pool. However, this research has shown, under the conditions studied, that the subset average discrimination may be used to adjust subset ability estimates to gain this equivalency. Similarly, if two or more groups of examinees are given different subsets of items, the expectation is that these groups ability estimates will not be on the same scale. Here again, scale comparability may be obtained by the use of average discrimination, and comparisons of the groups with respect to ability could be made.

The extension of this reasoning reaches its conclusion when each examinee responds to a different subset of items, i. e., tailored testing. In such a situation, if we record, and then average the item discrimination for each examinee's subset, we should be able to improve the ability estimates. This improvement is, again, in the sense of providing scale comprability for all examinee ability estimates.

Another usage of average item discrimination would be in the area of test equating. In utilizing the Rasch model to equate the tests of the Anchor Test Study, Rentz and Bashaw (1975, vol. 1) utilized two equating procedures. Their recommended procedure was called the "difficulty method" and is accomplished as follows:

> the difficulty method is based upon a Rasch analysis of both tests combined into one long test. If the two tests in the pair have the same scale origin, then the averages for log easinesses for both tests will be equal. To the degree that these averages are not equal, the two tests have scale origin differences. The equating constant is the difference

> in average log easiness. This constant is added to the abilities generated by one of the tests to put this test on the same scale as the other test. (Rentz & Bashaw, 1975, vol. 1, p. 50)

Under this method we may view the two combined tests as a total item pool, and view each separate test as a subset of items from this pool. The combined tests will have, under a Rasch analysis, an average item discrimination of unity. However, the separate tests, treated as subsets, may well have average discriminations that differ from unity. Therefore, the ability estimates from these two tests would be on different scales. Thus, one addition to the difficulty method of equating would be, prior to adding the equating constant, that of rescaling the ability estimates by the average discrimination of the test which produced the estimates.

Potential Uses of the Minimum Standard Error of Measurement

Based upon the standard error of an ability estimate, equation (1), it was shown that the minimum standard error is produced when all item difficulties match the ability of the examinee being measured. This leads us to note that, unless the probabilities for success on an item are extreme, the majority of items will contribute to the measurement of the examinee. However, the items with the highest contribution will be those which yield probabilities of .5. The implication of this is that we may judge effective test length by the number of items with probabilities near .5. Probabilities "near .5" could be evaluated as those items on the interval of, say, .4 to .6, or as the situation dictates. That is, when we desire a high degree of accuracy of measurement, we would desire item probabilities to be near .5. However, for quick, initial screening as to ability, we could relax our requirements and choose items with probabilities between, say, .3 and .7.

These concepts may be utilized in criterion referenced testing situations. In criterion references testing, the attempt is to classify examinees as to mastery or non-mastery. This may be accomplished by assessing whether an examinee's ability is above or below the ability level associated with mastery. This suggests selecting items whose probabilities are near .5 for the mastery level of ability. Thus our most accurate classifications would be at this point on the ability scale.

REFERENCES

Birnbaum, A. Some latent trait models and their use in inferring an examinee's ability. Part 5 of F. M. Lord & M. R. Novick, Statistical theories of mental test scores. Reading, Mass: Addison-Wesley, 1968.

Bouver, H. & Bargmann, R. Tables of the standardized percentage points of the Pearsonian system of curves in terms of β_1 and β_2 (Themis Technical Report No. 32). Athens, Ga.: University of Georgia Department of Statistics and Computer Science, June 1974.

Brink, N. E. Effect of item discrimination and the range of item easiness estimates using the Rasch model. Doctoral dissertation, University of California at Los Angeles, 1970. (University Microfilm No. 71-4863)

Brooks, R. D. An empirical investigation of the Rasch ratio-scale model for item-difficulty indexes. Doctoral dissertation, State University of Iowa, 1964. (University Microfilms No. 65-434)

Cartledge, C. M. A comparison of equipercentile and Rasch equating methodologies. Unpublished doctoral dissertation, University of Georgia, 1974.

Curry, A. R. Invariance of the Rasch Model Ability Parameter Estimates Over Different Collections of Items. Unpublished doctoral dissertation, University of Georgia, 1977.

Ferguson, G. A. Statistical analysis in psychology and education (2nd ed.). New York: McGraw-Hill, 1959.

Hambleton, R. K. An empirical investigation of the Rasch test theory model. Doctoral dissertation, University of Toronto, 1969. Available from R. K. Hambleton, University of Massachusetts, Amherst, Mass.)

Hambleton, R. K. & Cook, L. Introduction to latent trait models and their use in the analysis of educational data. Journal of Educational Measurement, 1977, 14, 75-96.

Hambleton, R. K., Swaminathan, H., Cook, L., Eignor, D. R. & Gifford, J. A. Developments in latent trait theory: A review of models, technical issues and applications. Paper presented at the annual meeting of the American Educational Research Association, New York, April 1977.

Hambleton, R. K. & Traub, R. E. Analysis of empirical data using two logistic latent trait models. British Journal of Mathematical and Statistical Psychology, 1973, 26, 195-211.

Hays, W. L. & Winkler, R. L. Statistics: Probability, inference and decision. New York: Holt, Rinehart and Winston, 1971.

Lord, F. M. An analysis of the Verbal Scholastic Aptitude Test using Birnbaum's three-parameter logistic model. Educational and Psychological Measurement, 1968, 28, 989-1020.

Loret, P. G., Seder, A., Bianchini, J. C. & Vale, C. A. Acher Test Study: Equivalence and norms tables for selected reading achievement tests (Grades 4,5,6). Washington: U. S. Department of Health, Education, and Welfare, U. S. Office of Education, 1974.

Panchepakesan, J. The simple logistic model and mental measurement. Doctoral dissertation, University of Chicago, 1969. (Available from University of Chicago Library.)

Rasch, G. Probabilities models for some intelligence and attainment tests. Copenhagen: Danish Institute for Educational Research, 1960.

Rentz, C. C. An investigation of the invariance properties of the Rasch model parameter estimates. Unpublished doctoral dissertation, University of Georgia, 1975.

Rentz, R. R. Evaluating the fit of tests to the Rasch model. Paper presented at the annual meeting of the American Educational Research Association, Chicago, April 1974.

Rentz, R. R., & Bashaw, W. L. Equating reading tests with the Rasch model (Report of U.S.O.E. Project No. OEC-0-72-5236, Vol. 2). Athens, Ga.: University of Georgia Educational Research Laboratory, September, 1975.

Rentz, R. R. & Bashaw, W. L. The National Reference Scale for reading: An application of the Rasch model. Journal of Educational Measurement, 1977, 14, 161-180.

Ross, J. An empirical study of a logistic test model. Psychometrika, 1966, 31, 325-340.

Schmidt, W. H. Necessity of the model. Paper presented at an American Educational Research Association pre-session on person-free item calibration and item-free person measurement, Minneapolis, March 1970.

Tinsley, H. E. A. An investigation of the Rasch simple logistic model for tests of intelligence or attainment. Doctoral dissertation, University of Minnesota, 1971. (University Microfilm, No. 72-14 378).

Whitely, S. E. & Dawis, R. V. The nature of objectivity with the Rasch model. Journal of Educational Measurement, 1974, 11, 163-178.

Willmott, A. S. & Fowles, D. E. The objective interpretation of test performance: The Rasch model applied. Unpublished manuscript, National Foundation for Education Research, United Kingdom, March 1973.

Wright, B. D. Sample-free test calibration and person measurement. In Proceedings of the 1967 Invitational Conference in Testing Problems. Princeton, N. J.: Educational Testing Service, 1968.

Wright, B. D. & Douglas, G. A. Best test design and self-tailored testing (Research Memorandum No. 19). Chicago: The University of Chicago Department of Education Statistical Laboratory, June 1975.

Wright, B. D. & Panchapakesan, N. A procedure for sample-free item analysis. Educational and Psychological Measurement, 1969, 29, 23-48.

APPENDIX A

Interrelationship of Model Parameters of the Normal Ogive, Birnbaum, and Rasch Models

It is possible that the lack of concensus regarding the robustness of latent trait models is due, at least in part, to a lack of knowledge concerning how the models interrelate. This problem is exacerbated by the use of non-standard terminology. For example, many researchers (usually those studying the Rasch model) refer to item slope rather than item discrimination. While ICC slope and item discrimination are related, their numerical values are quite different, thus much confusion arises. This section is provided in an attempt to alleviate these problems.

The Birnbaum or two-parameter logistic model is given by

$$P_{ij} = \frac{e^{Da_i(\theta_j - d_i)}}{1 + e^{Da_i(\theta_j - d_i)}} \text{ , where} \qquad (6)$$

θ_j is the ability of person j,

d_i is the difficulty of item i,

a_i is the item discrimination, and

D is a constant usually taken to be 1.7.

This model, unlike the Rasch model, allows for differences in item discrimination. The constant, D, is a scaling factor which is used to bring conformity between this model and the normal ogive model (Birnbaum, 1968, p. 399).

Birnbaum (1968, p. 402) presents the Rasch model, see below, as a special case of his model in which D = 1.0 and all item discrimi-

---

The Rasch model may be represented as

$P_{ij} = \exp(\theta_j - d_i)/1 + \exp(\theta_j - d_i)$, where

θ_j and d_i are defined as above.

nations are equal. The result of setting D to unity is that of employing a "pure" logistic model and not a model transformed to gain conformity with the normal ogive model. However, we may use D = 1.7 in the Rasch model to gain this conformity. This is allowed as the Rasch model makes no assumptions about the distribution of θ. The Rasch ability values, θ^r, divided by 1.7 will produce ability values on the normal ogive scale, θ^n_j i.e.

$$\theta^r/1.7 = \theta^n. \tag{7}$$

Having produced this change in the θ^r scale, it would be useful to ascertain the effects of this change on the Rasch model ICC. Since the slope of the normal ogive ICC, β^n, at the flex point is

$$\beta^n_i = a^n_i/\sqrt{2\pi} \tag{8}$$

the transformed Rasch ICC should have this same slope. This implies that we should be able to relate the untransformed Rasch slope, β^r to β^n.

Since $\theta^r/1.7 = \theta^n$, we can see that Rasch logistic scale is "expanded" when compared to the normal ogive scale. Therefore, when we "compress" θ^r by 1.7 we are raising β^r to make it equal to β^n. That is, since the slope is the ratio of rise to run, if we "reduce" run, we lower this ratio by a factor equal to the reduction factor. Thus

$$1.7\,\beta^r_i = \beta^n_i \quad . \tag{9}$$

Panchepakasan (1969, p. 41) gives the slope of the Rasch ICC as

$$\beta^r_i = a^r_i/4. \tag{10}$$

Since we know the relationship between β^r and β^n, we ought to be able to obtain the relationship between item discrimination under the

Rasch model, a_i^r, and normal ogive item discrimination, a_i^n. From equation (10) we obtain

$$a_i^r = 4\beta_i^r , \tag{11}$$

and from equation (8) we obtain

$$a_i^n = \beta_i^n \sqrt{2\pi} . \tag{12}$$

Noting the relationship in equation (9), we see that

$$(4)(1.7)\beta_i^r = \beta_i^n \sqrt{2\pi} .$$

Since $4\beta_i^r = a_i^r$ and $\beta_i^n \sqrt{2\pi} = a_i^n$, we may simplify (13) as

$$1.7a_i^r = a_i^n \tag{14}$$

Thus we see how the values of the Rasch model item discrimination are related to item discrimination values under the normal ogive model. Equation (14) provides a method of comparing discrimination estimates obtained under these two models. Furthermore, since the Birnbaum model employs the constant D, item discrimination under this model will be equivalent to the normal ogive discrimination.

The above relationships are based upon knowledge of either the slope of the ICC (at the point where $\theta = d_i$) or the value of item discrimination. If neither are known, they may be obtained through calculus. Alternately, we may have available the item and ability biserial correlations. Under the assumption of normality, this biserial may be obtained through

$$a_i^n = \rho_{bis} / \sqrt{1 - \rho_{bis}^2} . \tag{15}$$

Combining equations (14) and (15) we obtain

$$1.7\, a_i^r = \rho_{bis} / \sqrt{1 - \rho_{bis}^2} . \tag{16}$$

Therefore, if we have normal ogive data and wish to obtain the com-

parable Rasch model value of item discrimination we may solve (16) for a_i^r as

$$a_i^r = (\rho_{bis}/\sqrt{1 - \rho_{bis}^2})/1.7. \quad (17)$$

The above equations provide the basic points for interrelating the three most commonly encountered latent trait models. However, one additional point needs clarification.

A review of Rasch model research shows a common usage of the term item slope rather than item discrimination. The use of item slope stems from Rasch's research which was conducted independently of latent trait theory. In the development of his model, Rasch relied upon, not item characteristic curves, but upon the regression of "item log odds" on "test log odds." As noted earlier, the ICC is based upon the regression of item probability of success on ability. Item log odds may be obtained from the item probability by

$$\ln O_{ij} = \ln (P_{ij}/1 - P_{ij}). \quad (18)$$

Test log odds is actually the average odds for success on all items in the test. This average is equivalent to ability. Thus one major difference between the two regressions is the change from item probability to item log odds. The other difference is that the "log odds" regression produces a straight line graph. Rasch required that these log odds plots produce points which cluster around a line of unit slope (1960, p. 172). Thus the concern among Rasch researchers with item slope values.

Why does Rasch require lines of unit slope? The answer is found in the philosophy or goals used to develop the model. Rasch desired to produce measurements so that statements such as "person A is twice as able as person B" would have meaning. This implies that under the model, for an item of given difficulty, person A's odds for success

in that item would be twice person B's odds. If this relationship is "more than an ad hoc description of a limited set of data", the relationship should be in force in a more general context (Rasch, 1960, p. 9). Therefore, if we average over items in a test (items measuring a common trait), the relationship should be maintained; person A's average odds will be two times the average odds for person B. In plotting the transformed (by logs) item odds vs transformed average odds, the proper relationship will be depicted only if the plotted points fall along a straight line. Furthermore, since the two axes share a common metric, log odds, the slope of the line must be unity.

The question now becomes one of determining the relationship between the "log odds slope" and the ICC slope and discrimination. According to Panchapakesan (1969, p. 27), in the estimation of model parameters the origin of the difficulty scale is arbitrary. She further states that this origin is commonly set (to zero) by requiring the difficulties to average to zero. Also in Rasch analysis, the average item discrimination is set to unity with this information, let's consider the average item in a test, an item with a difficulty of zero and a discrimination of unity.

The plot of the regression of item log odds on test, average, log odds extends on both axes from minus infinity to infinity. Furthermore since this regression plot must intersect a line perpendicular to the zero point on the test log odds axis (because difficulty is zero), the line connecting the plotted points will be of unit slope. Thus, the value of the log odds slope is the value of item discrimination.

A FORTRAN PROGRAM FOR THE SIMULATION OF ITEM RESPONSES


```
C       A GENERAL PURPOSE DATA SIMULATOR FOR RESPONSES TO TEST ITEMS
C            ORIGINAL PROGRAM DEVELOPED BY:
                        R. ROBERT RENTZ - EDU.RSCH.LAB.-UGA
C
C       PROBLEM SPECS
C            NK      NUMBER OF ITEMS (MAX = 80)
C            ES      ITEM DIFFICULTIES
C            BTA     ITEM SLOPES (DISCRIMINATION)
C            ABIL    TRUE ABILITIES
C            NJ      NUMBER OF EXAMINEES
C            AB      RASCH MODEL ABILITY ESTIMATES
C            IDA     0/1 SIMULATED ITEM RESPONSE VECTOR
C            IXT     SEEDS FOR RANDOM NUMBER GENERATOR
C            FMT     ALTERNATE OUTPUT FORMAT
        DIMINSION  ES(36,80), BTA(36,80), AB(80), NJ(100),JCODE(36)
        DIMINSION  IDA(300,80), ABIL(3,100), ITX(36), NABILT(3),
        DIMINSION  KCODE(3), FMT(60)
        DATA       NJ/100*1/
        DATA       KCODE/1,2,3/
        DATA       NABILT/3*100/
        READ(5,490) KGRUP, FMT
  490   FORMAT(I1/20A4/20A4/20A4)
C       KGRUP = EXAMINEE GP CODE (1=NORMAL, 2=POS.SKEW, 3=NEG.SKEW)
        READ (5,550)  ICODE
  550   FORMAT(26I3/10I3)
C         ICODE=  ITEM SUBSET CODE NUMBERS
        DO 5551  K=1,3
 5551   READ(5,551) (ABIL(K,J), J=1,100)
  551   FORMAT(10X, 9(10F6.3/), 10X,10F6.3)
        DO 5552 K=1,36
 5552   READ(5,552) I , (ES(I,J), BTA(I,J), J=1,40)
  552   FORMAT(I2,18F4.3/3(2X, 18F4.3/))
        READ (5,553) NK, NOPT, ITX
  553   FORMAT (2I2,1X, 12I6/1X,13I6,1X,11I6)
        IGP = KGRUP
        DO 99 ISET = 1,36
C       PERFORM DATA SIMULATION FOR 36 ITEM SUBSETS
        DO 2 I=1,NK
    2   AB(I) = 0.
        IX = ITX(ISET)
        INUMB=NABILT(IGP)
        NSG=NK - 1
        IF(NOPT - 6) 3,3,4
    3   NOPT = 6
    4   ICOUNT - 0
C
C       STATEMENTS THRU 23 ARE FOR GENERATION OF
C       ABILITY VECTOR BASED UPON RASCH MODEL
C
        SC = .00001
        DO 23 J = 1, NSG
        DO 22 IT = 1,50
```

```
      G = -J
      GP = 0.
      DO 21 I = 1,NK
      XP = EXP(AB(J) - ES(ISET,I))
      PP = 1. + XP
      G = G + XP/PP
      GP = GP + XP/PP**2
   21 CONTINUE
      G = -G/GP
      AB(J) = AB(J) + G
      SD = G/AB(J)
      IF(SD**2.LE.SC)  GO TO 23
   22 CONTINUE
   23 CONTINUE
      WRITE (6,205) NK,IX, NOPT
  205 FORMAT( '1','THE ITEM RESPONSE SIMULATOR',
     A 12X, 'EDUCATIONAL RESEARCH LABORATORY'/
     B 2X, 'THE UNIVERSITY OF GEORGIA'/
     C 2X, 'R.R.RENTZ - DECEMBER 1974'///
     D 2X, 'NUMBER OF ITEMS=', I3/2X, 'RN STARTER = ', I61,2X,OUTPUT
     E UNIT = ',I2)
      WRITE (6,206) (ES(ISET,I), I = 1,NK
  206 FORMAT ('OGENERATING PARAMETERS - ITEM DIFFICULTY'//8(2X,10F8.3/))
      WRITE (6,207)(ES(ISET,I),I = 1,NK
  207 FORMAT('OGENERATING PARAMETERS - ITEM SLOPE'//8(2X,10F8.3/))
      WRITE (6,204)AB
  204 FORMAT('OABILITY GENERATED BY RASCH MODEL'//8(2X,10F8.3/))
C
C     DATA MATRIX TO BE COMPUTED BY EACH SCORE GROUP
C     INDEXED BY J
C
      DO 50 J = 1, INUMB
      N = NJ(J)
      IF(N.EQ.0) GO TO 50
C
C     COMPUTE PROB ACCORDING TO LOGISTIC 2 PARAMETER MODEL
C
      DO 10 I = 1,NK
      XLO = EXP(BTA(ISET,J)*(ABIL(IGP,J) - ES(ISET,J)))
      P = XLO/(1. + XLO)
      DO 7  K = 1,N
      IDA (K,J) = 0
C
C     FOLLOWING STATEMENTS FROM SSP SUBROUTINE RANDU
C
      IY = IX * 65539
      IF(IY) 5,6,6
    5 IY= IY + 2147483647 + 1
    6 YFL = IY
      YFL = YFL * .4656613 E-9
      IX = IY
```

```
C
    7  IF(YFL .LE. P) IDA(K,I) = 1
   10  CONTINUE
       DO 12 K = 1,N
       ICOUNT = ICOUNT + 1
       KRS = 0
   11  KRS = KRS + IDA(K,I)
       KKS = KRS
       IF (KKS.EQ.0) KKS=1
       IF (KKS.EQ.NK) KKS = NK-1
       ABDIFF = ABS(AK(KKS) - ABIL(IGP,J))
C
C      ABDIFF = ABSOLUTE DIFF OF RASCH MODEL ABIL. ESTIMATE
         MINUS TRUE ABILITY
C
       IF( K.EQ.1. AND. NOPT.NE.6) GO TO 20
       GO TO 12
   20  WRITE (6,2010) KCODE(IGP),ICOUNT,ABIL(IGP,J),ICODE(ISET),
     A KRS,AB(KKS),ABDIFF,(IDA(K,I), I = 1,NK)
 2010  FORMAT(' ',I2,I5,F7.3,I3,I5,F8.4,F9.4,10X,80I1)
   12  WRITE(NOPT,FMT)KCODE(IGP),ICOUNT,ABIL(IGP,J)
     A ICODE(ISET),KRS,AB(KKS),ABDIFF, (IDA(K,I),I=1,80)
   50  CONTINUE
   61  WRITE(6,202)ICOUNT,NOPT
  202  FORMAT('0','THIS PROBLEM COMPLETED EXECUTION AND WROTE ',I6,
     A 'RECORDS ON UNIT', I3)
   99  CONTINUE
       STOP
       END
```

APPENDIX C

ITEM SUBSET DESCRIPTIVE STATISTICS

Descriptive Statistics for Easy Subsets

Difficulty

$\bar{X}$ = -1.000 s.d. = .468

Subset	Discrimination					Random Number
	$\bar{X}$	s.d.	min.	max.	r(Diff, Disc)	Seed
1 (0%, Low)	.997	.392	.502	1.498	.197	896993
4 (0%, High)	.999	.639	250	1.750	.066	563162
7 (0%, Mixed)	.985	.517	.250	1.690	.064	213476
10 (25%, Low)	.990	.342	.502	1.486	.125	907309
13 (25%, High)	.974	.564	.250	1.738	-.249	173635
16 (25%, Mixed)	1.011	.460	.274	1.702	-.024	225391
19 (50%, Low)	1.027	.290	.538	1.498	.177	733531
22 (50%, High)	1.007	.471	.262	1.726	.096	647873
25 (50%, Mixed)	1.021	.396	.298	1.750	.074	674700
28 (75%, Low)	.998	.224	.550	1.462	-.137	742287
31 (75%, High)	.993	.364	.250	1.738	.096	318708
34 (75%, Mixed)	.995	.333	.274	1.690	-.127	025492

Descriptive Statistics for Hard Subsets

Difficulty

$\bar{X}$ = 1.000 s.d. = .468

Subset	Discrimination					Random Number
	$\bar{X}$	s.d.	min.	max	r(Diff,Disc.)	Seed
3 (0%, Low)	.997	.386	.502	1.436	.147	061696
6 (0%, High)	1.006	.642	.250	1.750	-.196	612758
9 (0%, Mixed)	1.002	.522	.262	1.738	-.113	219982
12 (25%, Low)	.986	.347	.520	1.498	.153	757940
15 (25%, High)	.998	.564	.250	1.738	-.168	502942
18 (25%, Mixed)	1.000	.459	.250	1.738	-.121	057957
21 (50%, Low)	.995	.238	.526	1.474	-.001	385187
24 (50%, High)	.997	.459	.250	1.726	-.157	674707
27 (50%, Mixed)	1.019	.408	.250	1.726	-.074	216423
30 (75%, Low)	1.008	.239	.538	1.438	.080	827594
33 (75%, High)	.998	.354	.274	1.702	.211	585045
36 (75%, Mixed)	1.001	.299	.418	1.694	.004	131370

Descriptive Statistics for Moderate Subsets

Difficulty

$\bar{X}$ = 0.000 s.d. = .468

Subset	$\bar{X}$	s.d.	Discrimination min.	Discrimination max.	r(Diff,Disc.)	Random Number Seed
2 (0%, Low)	.993	.395	.574	1.498	-.091	561744
5 (0%, High)	.994	.642	.250	1.738	-.143	497917
8 (0%, Mixed)	.998	.524	.274	1.702	-.007	956567
11 (25%, Low)	.986	.370	.502	1.498	-.073	354379
14 (25%, High)	1.009	.561	.252	1.738	.162	083343
17 (25%, Mixed)	1.008	.451	.250	1.678	-.255	729125
20 (50%, Low)	1.024	.291	.538	1.498	-.147	091281
23 (50%, High)	1.001	.456	.252	1.714	.047	896857
26 (50%, Mixed)	.994	.390	.250	1.714	.202	238139
29 (75%, Low)	995	.254	.526	1.498	.142	423321
32 (75%, High)	.992	.328	.250	1.582	-.090	580614
35 (75%, Mixed)	.987	.282	.298	1.594	.034	593462